The Dissertation Committee for Brian Arnold Soeder Certifies that this is the approved version of the following dissertation:

A Trust Based Methodology for Determining Identity Risk

Committee:

K. Suzanne Barber Supervisor

Aristotle Araposthatis

Vijay Garg

Robert Flake

Matthew McGlone

Robert Metcalfe

A Trust Based Methodology for Determining Identity Risk

by

Brian Arnold Soeder, BS, MSE

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

May 2015

Dedication

Dedicated to my wife Natalie for her unending support of my research and career.

Also a special dedication to my son Shane who was born during the process.

Acknowledgements

I would like to express my appreciation for all of my colleagues and professors. My company, the MITRE Corporation has given me the time and freedom to execute my research. I would like to convey my gratitude to the entire family at the Program Executive Office for Command, Control and Communications Tactical, especially PM Mission Command for their patience as I finished my dissertation.

I would like to thank the team in the Center for Identity. I'd like to thank Suratna Budalakoti, Dave DeAngelis, and my advisor, Dr. Suzanne Barber for her confidence and trust. A special thanks to Elizabeth and Andy for their help in editing and shaping concepts. Further, I'd like to thank my committee members Dr's Metcalfe, Flake, Garg, Araposthatis, and McGlone for all of their guidance and assistance in completing the research.

I would like to express a personal thanks to Rob, Sam, and Tim for their encouragement, support and logistics. Finally I would like to thank my parents, wife, son and dogs for their constant support and patience.

A Trust Based Methodology for Determining Identity Risk

Brian Arnold Soeder, Ph.D

The University of Texas at Austin, 2015

Supervisor: K. Suzanne Barber

Identity theft, fraud, and abuse are rapidly increasing around the world, yet how information systems security providers verify user credentials remains generally unchanged[32]. In order to grant users access to services such as bank accounts or social networks, providers must collect information to store as user credentials. Previous research using authentication and authorization approaches has examined validating user credentials and controlling access, but these approaches still fall short in accurately *identifying* users[48]. To confirm user identity, most protocols only offer a binary indicator. Thus, quantifying the levels of trust between service providers and their users is necessary but not sufficient for ensuring secure transactions. In the context of transactions, this research proposes leveraging credential attributes to improve confidence in a user's identity leveraging trust and risk management methodologies.

Transactions between users on the Internet require credentials that have a fixed number of attributes. When these credentials are created, attributes such as Social Security number, mother's maiden name, and address are used to validate a particular user. Attributes are often lost, stolen or compromised. Once the attributes of an identity are compromised, anyone can assume that identity with benign or malicious purposes. Traditional solutions to this problem are to increase the trust level of the authentication through multiple modes, such as biometrics or smartcard tokens. While biometrics and

smartcards are very useful attributes for increasing trust, this research shows that it is possible to increase trust of users with attributes typically held by or known to the user. This approach is appropriate in terms of cost and convenience and scales to a large number of transactions. Using only the attributes registered with an identity provider (e.g., address, zip code, name, etc.) can show how trusted a user is who presents an identity. Further, the risk to a service provider of allowing access to that user can be established with this limited information.

Specifically, this research approach correlates attributes with existing information, including patterns of using attributes to authenticate the user and trustworthiness of existing data maintained by identity providers. The ability to correlate and vary these attributes provides higher confidence in a presented credential. The proposed methodology is shown experimentally to more accurately assess the risks of granting users access to a given set of data and information then existing approaches. These approaches to identification are shown to significantly increase the confidence of credentials granted to individuals through a series of simulations representing common transactions involving identity.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Suppose that on Monday we cast a certain bar of metal into a statue. Then on Tuesday, we melt the statue down and recast the metal into a vase. And on Wednesday, we melt the vase and are left with just the piece of metal. Surely the statue was the piece of metal on Monday and the vase was the piece of metal on Tuesday. But the vase was not the statue and neither one of these was the piece of metal on Wednesday.

Roderick Chisholm (1973)

The above quote highlights the current state of online identities. The statue and the vase are composed of the same material, yet the true identity of either object cannot be detected solely by looking at their compositions. Similarly, vast amounts of online information compose human and system identities, and this information is stored in different locations all over the world. Yet, we cannot ascertain the true nature of an identity simply by examining the composition of this information. Understanding how to untangle these complex sets of information is a challenging task, but it is possible to leverage existing online information about people and systems to gain confidence in their identities.

Businesses see significant amounts of consumer identity fraud and have little ability to react to it. Identity fraud and abuse is estimated to cost businesses billions annually [44]. Massive data and information breaches are now commonplace and compromise users' identity attributes (e.g., passwords and Social Security numbers). By compromising these critical pieces of data, thieves can also compromise other third-party systems where users' credentials are stored [42, 43]. The financial burdens of compromised identities have grown so great that the U.S. government formed the National Strategy on Trusted Identities in Cyberspace (NSTIC). NSTIC was launched to raise identity trustworthiness in online transactions by enlisting a team of public and

private entities. In order to combat common points of failure identified by the NSTIC and others, this research proposes an approach that dynamically measures user reputation and experience to build greater confidence in user identity. Specifically, using the context of online transactions is one way to examine the different types of trust required between people and systems during online identity creation and verification processes. Ultimately, the risks associated with trusting identities are also quantified. This research delivers new methods and algorithms that provide:

- Quantifiable measure of trust in an identity;
- Increased precision of trust in an identity using less expensive information-based authentication;
- Predict the information on which to rely when authenticating an identity.
- Assess risk for service providers before granting access to information

1.1 IMPORTANT DEFINITIONS

In order to more concretely define the scope of this research, the following conceptual definitions are used throughout the dissertation. First, in accordance with the Personal Identity view [57], an ***identity*** (I_p) can be defined as an almost infinite list of ***attributes*** existing within a person $I_{p=}\{A_0,A_1 \ldots A_i\}$. These attributes consist of personally identifiable information (PII) about an individual such as name, address, and Social Security number. An identity can apply to any system, person, or thing composed of parts which are inherently inseparable from the whole. An identity is also unique such that person or a system only can have one. A ***user*** is defined as an individual using an identity. All online transactions begin with the enrollment process, where a user presents different attributes to create ***credentials*** such as email addresses and passwords with an identity provider or service provider. It is important to note that individuals can use one

or more identity credentials simultaneously [1, 13, 57]. Therefore credential $c \subseteq I_p$, where $c = \{A_0, A_1, \cdots, A_j\} \; \forall \; A_j \in I_p$.

Figure 1 represents an example identity as modeled in this research. The identity is composed of a credential and multiple attributes with specific relationships to both the credential and the identity. The identity value I_0 represents the index in a connected graph model to be discussed later.

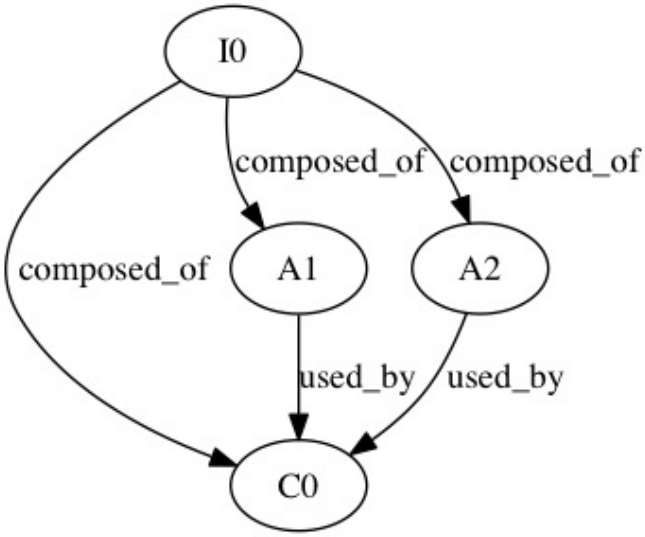

Figure 1: Graphical Model of Identities, Credentials, and Attributes

Each node in the graph can be related to a corresponding type in the Identity Ecosystem [18] that will be discussed in further chapters.

In this research, the terms identity provider and service provider have distinct meanings but sound the same.. For example, a user could enroll with Facebook, which would be considered a service provider, by presenting different identity attributes. Facebook verifies the presented attributes and the user becomes an authorized user with credentials (e.g., user name and password). The same user could also use their Facebook

credentials to log into another service, such as Twitter. In this situation, Facebook would be an identity provider, and Twitter would be a service provider.

This overlap also points to a common problem in identity security. Identity verification processes vary greatly by identity provider and are not always transparently communicated between different providers. This can create an environment of distrust, even though users generally present the same set of attributes when enrolling with different providers [22]. Credentials can be created from the same identity using different methods and combinations of attributes depending on the decisions made at enrollment time. Different companies and agencies can also store, govern, and manage credentials using different rule sets. For example, the magnitude of available user credentials with different providers in 2012 is illustrated in Table 1.

Credential Category	**Number in 2012 (millions)**
Smartcard	6925
Social Media top 10	2380
Email Sites top 10	2066
Internet User Current Population	2267
Credentials Per User	**At least 5.01**

Table 1: User Accounts on the Internet [46, 47, 51, 52]

Table 1 establishes with some degree of certainty that any individual Internet user has, on average, at least five credentials. By obtaining only a few of those attributes, a thief could compromise the information associated with more than five of a user's credentials. This overlap in attributes could also be causal to any number of identity-based threats to the user. If identity providers converge on a set of attributes that are slow to change, then the probability of compromise in those attributes grows, then the likelihood of identity theft could decrease [4].

Convergence is one of the driving factors of the Federal Identity Credential, and Access Management (FICAM) Implementation guidelines [30]. These guidelines discuss the close relationship between three important processes of identity creation: credentialing, authentication, and authorization. These processes verify whether or not an identity is trustworthy based on available attribute information and are used for access control decisions [4, 9, 13]. In other words, these processes demonstrate implicit trust between the users and identity providers, and usually result in a "yes" (e.g., this user or provider *is* who they are claiming to be) or "no" (e.g., this user or provider *is not* who they are claiming to be).

The lack of variety of attributes used also adds to difficultly in trusting identity providers. In this research, ***trust*** is a quantifiable factor that captures the degree of confidence maintained in a piece, or collection of data [15]. Trust is also directional between the service or identity provider and the user. This research focuses directionally on the trust of a user's identity by an identity provider. By simply ensuring a sound distribution of attributes and varying their usage, identity providers are more apt to be trustworthy. Unfortunately, the lack of variety in attributes, slow rate of change, and large overlap of attributes between identity providers leads to higher threat vectors in identity compromise.

Identity providers offer confidentiality, authentication and authorization services to a variety of users. These providers can be compromised by other types of simple attribute theft from other Identity providers [49] and not even know it! Trusted repositories are being compromised by a variety of vectors such as DigiNotar, Gemnet, and others [34]. Operating system and browser providers also maintain lists of trusted identity providers based on less-than-explicit criteria [50]. In this type of marketplace, users have to either pay for a certain type of key (i.e., a key with an *apriori* level of trust

set by the application vendor), or use some other type model for credentialing. This problem is magnified for the service providers who have to determine which types of credentials to accept. There is still no formal, widely accepted mechanism for determining a quantitative value of trust placed in these repositories.

1.1.1 Existing Authentication Solutions

While there are significant challenges with authentication on the Internet, there are different categories in the solution space. These categories will be discussed more in section 2. One solution category is to add higher integrity to the enrollment process. This control strategy has lead to increased focus on authentication and authorization, the two main transaction protocols using identities. A common method to control the creation of identities is to conduct high assurance enrollment [40, 49]. This type of enrollment provides for escalating degrees of verification to include physical identity verification as well as biometric processes. Using a Public Key Infrastructure (PKI), identity providers can issue credentials and maintain a degree of certainty that the trust inherent in the infrastructure will provide cryptographic certainty during authentication. The authentication protocols then support a check to the infrastructure to ensure that the PKI is valid. Unfortunately, controlling enrollment usually creates more attributes and takes significantly more time. Often times, it adds more of a burden to the end user.

Another category for authentication is introducing multiple factors [37]. Asking users for something they know (e.g. their mother's maiden name), something they have (e.g. drivers license) and something they are (e.g. finger print scan) strengthens the authentication protocol, especially if these concepts are used in combination with one another. Similarly to the aforementioned distrust between identity providers based on enrollment processes, these stores of identities are disconnected and distrustful of each

other. Some researchers have tried to connect credential trust to provider trust during the authentication process [9, 24]. This connection is useful to an authentication transaction it sets the theoretical foundation for the usage of a probabilistic authentication process.

Finally the last category of solutions in authentication encryption is the most commonly used. An Identity provider can cryptographically validate an identity [50]. This method has two fundamental approaches. The first is to simply encrypt traffic from one the sender to the receiver. This encryption is limited to ensuring that no one can see the traffic as it transits the network. Encryption helps with confidentiality of the attributes inside of credentials but falls short on ensuring that the end user is the one who actually owns the identity. The other approach is the cryptographically sign an identity. This is similar to the Public Key approach discussed earlier [20]. It can ensure the sender of the credential has some ownership over his attributes. Signing of credentials falls short of making authentication easier because it requires some kind of an out of band issuance of these cryptographic credentials.

1.1.2 Existing Authorization Solutions

Other researchers have focused on solving these fundamental trust problems with access control protocols. A significant category for adding trust to transactions using access control protocols is based on role/attribute [10]. An identity store with some degree of trust can either supply information about the user through attributes directly or can do so in a federated manner. This approach allows more granular access policies to be built and ultimately increases the level of effectiveness in identification due to the abstraction layer that attributes and roles provide. The drawback to Attribute Based Access Control is the need to define a policy for each type of information that is used. An example would be a policy that *apriori* defines a user with a certain collection of

attributes will be allowed to access a piece of data. It fails to allow for access to others who need the data but are not high risk.

Another important factor of access control is that it provides an abstraction layer that a service provider can use to view users in the form of an identity rather then just a credential. Access control techniques have been based on trust as well. Liu [06] has taken the concept of trust derived during the authentication procedure and abstracted it for use during the access control transaction. This formalism provides a way to view authentication and access control in the same model, which is useful. When coupled with the more robust means to do access control, and the reuse of authentication transaction data, Snyder's view of Protection in Access control is immensely useful [11]. He abstracts the users to types of access and shows relationships with the user and the service through a graph. These graphs allow a query to answer specific questions about how to make dynamic access control decisions based on these relationships. This category of approaches falls short due to the formalism needed to conduct a transaction. Each relationship must be predefined and connected with the types of data to be controlled. This assumes that all data is known as well as all possible types of relationships. A more practical approach is needed that is responsive to service provider needs.

1.2 Problem Definition

Service providers must conduct a set of transactions to ask the authentication question, "who to the person/entity seeking the given service" and the authorization question, "is the person/entity eligible for the given service." Depending on the service, answering these questions may be very simple and occur simultaneously (e.g. access to a sales event site of your favorite retailer) or be very complex and separate (e.g.

authentication and authorization at immigration). Well before the user transaction with the service provider occurs, the User U_1 enrolls with the identity provider for I_1 with a certain set of attributes. U_1 is granted a credential C_1 with which he can access a certain set of services. C_1 is associated with one identity provider and using only a subset of its attributes such as a username and password, the User can then present C_1 to service provider S_1. S_1 then passes these attributes to I_1 and if validated, the authentication transaction is complete. In other words if validated, I_1 asserts that the attributes from C_1 equal what is expected from U_1.

There are many situations in which this might not be the case. In the next step of the transaction, S_1 validates that C_1 has access based the set of attributes deemed necessary for access to specific data or resources. Again, the attributes from C_1 represent a probability of 1 that it represents U_1 although it is in an entirely different context from the Authentication transaction. Finally, U_2 can compromise the entire process by stealing the necessary attributes to represent itself as U_1 by recreating C_1. Therefore it is clear that the probability of C_1 representing U_1 isn't 1. It must be lower based on several factors. **This dissertation asks a series of questions that generates probabilities in each step of these transactions that represent the confidence that U_1 is represented by C_1.** Specifically, the dissertation assesses probabilities for the following steps:

1. **Attribute Compromise Probability**- As Attributes A_i are used it is important to evaluate their utility. Attributes are more useful if there is a probability that they haven't been compromised. Without an understanding of compromise possibility, Identity Providers could and do use the same attributes expecting a lot more utility from them.
2. **Authenticity of a User U_1**- As User U_1 conducts transactions, it is important that the attributes that he uses have values that are considered valid based off

of the identity repository. Authenticity is useful as a probability because it provides the ability for a user to have some invalid attributes but still provide evidence to an authentication decision. A user won't always remember everything but the more they know, the more authentic they are.

3. **Reliability of a User U_1-** In addition to getting attributes correct, the repetition of correctness of a specific attribute adds to the level of trust of the identity. Reliability is expressed as a probability that indicates the probability of the user providing consistently correct credentials. The utility of this metric is primarily the ability to describe whether an identity thief is using a stolen identity. If so the identity provider will have the ability to respond either by asking for more attributes or by some other means.

Taken together these probabilities form an aggregate confidence metric expressed as risk. Risk is the combination of each of these factors expressed as the probability that a user is who he says he is. Risk is useful as a single metric for all identity and service providers to reason about whether to authenticate or authorize a user.

1.3 RESEARCH QUESTIONS

Overall, this dissertation proposes the following hypothesis:

Using a confidence metric for transaction-based identity processes increases the accuracy of identification.

A confidence metric in an identity can be established by decoupling identities from credentials. After they are decoupled, a calculation of trust involving authenticity, reliability, and risk can be performed in order to represent an aggregate view of confidence in identity. Accuracy of identification in transactional systems is based on identity (i.e., thing or person) presenting an artifact represented as credential. Once the

credential is validated, the user can perform any action that is allowed from that credential. In practical usage of identity, transactions that use identities are typically very tightly coupled between authentication mechanisms and authorization mechanisms. For example, there is a one-to-one relationship between authentication of a user and authorization of that same user to most email systems used on the Internet. In order to peel apart the layers of identity usage, certain questions can help frame the problems this dissertation addresses. One important question that must be asked about an identity in a transaction is:

What evidence is there that the person or system that is using a credential is who they say they are?

There are many facets of this question. A user could be using a credential fraudulently. Users often share account passwords with other users; in this situation, a different user is clearly using a credential. Although the action is usually performed with benign intentions, the possibility illustrates a shortfall of current mechanisms. If one user enables another user to impersonate them, the impersonation is likely possible when using any credential short of strong biometric identification [29]. More often though, a credential is compromised by missing pieces of information that harm the identity, such as attributes that have changed over time, changes of users, and inherent issues with the protocols governing these attributes. Calculating the believability in an identity helps to overcome the changes that have happened post-enrollment and throughout the identity transactional lifecycle. Another important question to ask about an identity during a transaction is:

Is the user actually using credentials for malicious purposes during the transaction?

Assume an identity provider issues a credential to a user under the belief this identity is a benign actor with a legitimate purpose. Information can be derived about a

user's identity based on the actions it takes while using the credential. Using simple attributes of a credential during a transaction, it is possible to develop an identity reputation. For example, if a credential is used to access data that deviates from its reputation in a transaction, that credential and/or identity could be flagged. This confidence metric can represent the missing information to support an access control decision. The final question to ask about an identity during a transaction is:

Is the identity presenting a risk to information in the transaction?

This question is distinct from that of malicious purpose; an identity could also present direct risks to data in the transaction. A user's poor reputation, combined with the importance of the data access, can provide a third component to identity confidence.

Furthermore, the aforementioned hypothesis will be examined using three specific research questions about identification processes and identity confidence.

1.4.1 Research Question 1

RQ1: Which attributes of identity are most trusted?

Trust in identity attributes is a critical underpinning to any use of that Identity. Since an Identity is composed of these attributes, one must be able to reason about which ones are used, and compromised to have any sort of ability to define higher level trust or risk metrics. Today these attributes go through empirical evaluation methods to determine the most compromised [48]. In order to increase trust levels in a user, it is helpful to use the least compromised attributes to identify that user.

In order to answer this question, this research proposes first building a model to view the relationships between attributes and credentials. Credential repositories are initialized when users enroll with a certain set of attributes. Each provider is limited by its enrollment process and contained attributes. Attribute types can vary, but some are

more often used than others across identity providers. The connections of attributes form the basis for demonstrating the most compromised attributes [68]. A review of these attributes in use by identity providers will be conducted to bootstrap the initial data for a graph of attributes. After building the initial model of attributes to represent an identity, the connectedness of these attributes will be assessed following a social graph approach [85]. Centrality of nodes is then calculated using Eigenvector Centrality. The most central nodes in the graph become the most likely to be compromised. Probability of Fill is also introduced as a way to reason about how frequently attributes are used across credential repositories.

1.4.2 Research Question 2

RQ2: Can the reliability and authenticity of an identity in a transaction help demonstrate trust to identity?

An identity's reliability and authenticity in transactions are necessary to compute trust in that identity. Without establishing the user of the identity is using the correct attributes (authenticity) and that they are being used repeatedly (reliability), trust cannot be accomplished.

Building on the reference model of attribute characteristics established through RQ1, RQ2 aims to capture what available information at the start of a transaction. The answers to RQ1 help us understand which attributes of a credential are important for calculating confidence. The attributes collected in the transaction are compared to the golden records for identities, and the comparison yields a factor representing the trustworthiness of a credential's source based on the types of attributes that it uses. Other comparisons can be made between the values of each attribute. For example, if the

values of attributes in one credential supporting an identity do not match the values in another related credential, overall trust is diminished.

1.4.3 Research Question 3

RQ3: Is the risk posed by an identity a reliable predictor of the validity of the identity?

Asking the question about the risk posed by an identity helps a service provider decide whether to grant access or not. Returning this risk as a probability that a user is valid, has powerful implications. The simple yes or no approach to authorization can be replaced with a much more robust approach that incorporates information about the user and how they have behaved over time.

In RQ1 and RQ2, differing components can be calculated by relating their values. RQ3 compares multiple methods of combining trust and attribute compromise modalities. The first method is based on a weighted sum approach [27] and is called Improved Reliability. This approach allows the reliability metric to be directly related to the other the compromise probability of an attribute. Another method involves a future risk calculation using a Bayesian network approach [38, 39, 75]. This approach allows the risk calculation to make inferences about the potential future state of the Identity I_0 at a future time t_{1+x}.

Future state inferences are used to reduce computational complexity and normalize error at future states of the system. Results will be evaluated to determine if confidence in identity is added. The algorithm will reuse guidelines that consider the impact severity of disclosing information [19, 22]. When combined with calculated probabilistic risk, this approach provides a more traditional view of risk based on severity and probability of occurrence. The new risk model will be compared to more general

risk-based models [2, 19, 22]. Results will be evaluated for differences in risk levels between correctly used identities and incorrectly used (or downright false) identities.

Chapter 2: Related Work

2.1 Identification Methodologies

Existing research dealing with identifying a person or entity often involves third parties. With instances of identity fraud growing over 5 percent from 2011 to 2012, many users are concerned about the safety of their personally identifiable information [32]. One source of concern is supposedly trusted, yet oftentimes weak, third party identity providers. High profile attacks on Public Key Infrastructure (PKI) certificate authorities such as DigiNotar underscore this point. When PKI's are breached, many dire and fraud-related consequences to the end user can occur [32]. The very fact that these certificate authorities make credentials available over a network makes them attractive targets for attack [25]. Such ID repositories can have perfectly adequate methods for proofing, validation, management, dissemination, and storage, but still be hampered by basic side channel attacks based on infrastructure protection [25].

None of these systems have proven certain parties can be trusted to manage authentication or access control [5]. This flaw becomes a critical assumption of most researchers dealing with large client server networks. Many protocols including PGP and PKI have relied on trusted third parties to authenticate. In the web-based world, confidence in identity is a larger problem not typically dealt with, but authentication with a known repository is common. In even the most rigid of authentication schemes, certain bad actor vectors can defeat the practical application of authentication. Passwords can be stolen, biometric data can be compromised, and replay attacks can happen over insecure channels.

Thus, this research assumes that looking at authentication with a statistical approach is more useful. Authors writing authentication protocols make many

assumptions, and if one assumption is invalidated, then the entire protocol is at risk. The first potential flaw is that most authentication protocols make some assumption about the persistence of an authenticated user [7, 8]. In the transactional space of the web, regardless of the type of application layer configuration, each transaction is characterized by the atomicity of its operations, largely due to HTML and the static nature of the World Wide Web [9].

Another important assumption analyzed in this dissertation is that a trusted third party is becoming less viable. When assuming that a third party manages authentication, most previous research fails to account for the attributes of the third party, which could decrease trust in that third party. Trusted third parties must be reachable over networks they do not control. Since they can't control the path, these parties can't fully guarantee privacy. They also employee workers who have the passwords and keys to trusted infrastructure, which could potentially compromise transactions for varying purposes. Third parties can also have enemies that employ out-of-band and other techniques to compromise transactions for varying purposes. Based on these factors, this research questions the assumption that trusted third parties can even exist.

The final assumption this research makes that differs from common authentication models is that relationships between consumers and providers are dynamic. Most authentication protocols involve the pairwise concept of consumer and provider. A statistical approach demonstrates that a node acting as a consumer could be a republisher, and thus might potentially warrant less trust. For example, in current U.S. Department of Defense (DoD) networks, many intermediary sites download and republish data. These sites often masquerade digitally as end users by hardcoding in user names and passwords. Some sites also open up unnecessary back doors for reasons that aren't always fully understood by content owners until it is too late. A statistically based authentication

mechanism allows statistical inferences about where data goes that might help data owners to better understand and meet emerging requirements.

Beyond simple authentication lies the problem of trusting users with access to certain types of content. This problem is usually passed on to the authorization or access control mechanism to deal with. The fundamental assumption made by that component is that authentication has already taken place, whether through a trusted third party, local source, or other means. Most authorization mechanisms, such as Role Based Access Control (RBAC) have the ability to perform type authentication and determine permissions [10]. Authorization mechanisms also have detailed understandings of the nature of protected content. In some cases, the content could be reflected by a database in which each table, record, or column tagged with certain access permissions. In other cases, access control could be granted based upon combinations of attributes within a Web service. These cases simply depend on the protocol being used and the access mechanism chosen.

Given the atomic nature of transactions and authentication, combined with an assumed authorization, the risk of each transaction can grow. In order to quantify this risk to a system, we must relate the state of an authentication with the corresponding data a particular transaction is attempting to access. This relationship will help build more robust models and possibly even allow for direct feedback to the authentication system. At a minimum, the model should account for standard characteristics of the user to use for correlation.

Overall, identification of a person or entity is a problem that can be approached from a psychological, biological, or network-based perspective. Each view has the ability to contribute toward a common goal of identifying people or systems for specific transactions. The psychological or humanistic view of identity focuses on the qualitative

attributes of a person [26]. From this view, many factors can affect identity, including age, self-perception, role, and mental view [18]. Each one of these attributes can be viewed as properties of an individual's persona or credential, as we will later discuss. Typically, an online identity is associated with a persona, thus allowing an individual with a singular identity to have multiple personas. One way to model this view comes from Kini and Chooineh, who describe the relationship between humans and electronic commerce [24]. The derivation of a persona from a human identity involves an interpersonal view, a societal view, and a relational view. Other research in biometrics and health sciences incorporates a different perspective on identity.

Biometrics is the art and science of identity verification based on human characteristics [28]. Biometric identity verification results in accepting or rejecting the identity claim made by a given person based on biological attributes such as fingerprints and DNA [29]. The long-established goal of biometrics is making identification a fully–automatic process. Most biometric algorithms focus on a single modality to establish identity, and the most relevant aspects of automated identity verification to this research are multiple modality biometric systems. These systems demonstrate significant confidence improvement by fusing single modes using approaches such as summing [28, 29]. Along with confidence improvement, some challenges emerge when performing information fusion, including the required performance overhead [31]. False positives and false negatives are also important biometric approaches. These are measured based on the characteristics of the equipment used for each mode of identification. Most single modal algorithms assume each mode follows a Gaussian distribution [29]. This assumption leads to the same number of false positives and thus, a very linear distribution.

Multimodal approaches to identification can help to decrease the instances of false positives and negatives by providing a more robust data set [29]. These approaches can be combined by different weighting factors, which form different types of distributions. These usually form a non-parametric distribution, which keeps the evaluation from focusing on the dimensionality of the result. Using multiple modes of understanding also provides for greater fidelity and reliability overall [31].

All of these existing approaches to identification have their shortfalls. Some rely on finding a large amount of additional information attributes about a person to increase the strength of identification. Some can only identify parts of an identity that are not sufficient to ensure a consistent usage pattern. All of these methodologies fall short due to a need to produce a singular result in order to take an action on. This result must either be to accept or reject the identity. The utility of a probabilistic approach is the flexibility of use in determining whom to trust. It is possible to set a threshold based on the need for integrity and strength of identification.

2.2 TRUST METHODOLOGIES

Trust can be modeled from social, economic, computational, and cognitive perspectives [60]. Since trust characteristics are highly dependent on the system they are measured in, and identity is very uncertain in e-service systems, trust will be used in this research as a tool for increasing confidence in identity. The factors below contribute to formalization of trust in this dissertation:

- Trust is has a value between 0 and 1
- Trust is directional
- Trust is time dependent
- Multiple classes of trust factors are interdependent

With these factors, trust becomes a superset of the measures used to calculate a better confidence in the identity of a user. More specifically, this dissertation applies previous research on belief and reputation toward combining multiple modalities to form a better confidence metric [1, 3, 4, 60].

2.2.1 Belief

The initial condition of a connection to a service provider requires a set of beliefs about the user that initiated the connection. In most authentication protocols, that belief is based solely on the credential presented and the implied reputation of its source. Any sort of connection from a user involves implicit vulnerability [1]. Thus, it is important to build a better model of the user or identity that is attempting to connect. Gefen posits that any consumer belief is based on a three dimensional scale of trustworthiness including ability, integrity, and benevolence of the vendor [1]. These measures reflect the view of the user to the online service provider. Metrics from these dimensions can be used to establish confidence in the identity provider based on the credential presented.

Maurer presents another perspective of the belief of credentials that is based on a confidence parameter in the statement presented to an agent [20]. In this view, each statement earns specific confidence values based on the calculated probability in the truth of the statement. For example, a statement is a certificate credential, and the agent calculates trustworthiness based on a certification chain of the credential within a PKI [20]. Like Maurer, this research assumes that a third party identity provider cannot be fully trusted, but does have explicit trustworthiness based on certain factors. In the credential digraph from Maurer's example, each link in would make a belief statement about the previous node based on its authenticity as determined by different factors. In Maurer's model, these factors depend solely on the links between the different

certification nodes in a PKI when viewed as a graph. This reduces the belief question of a credential down to a calculation of probabilities based on vertices and edges in a graph.

While useful for helping to establish initial confidence in an identity, Maurer's view lacks some utility because of the computational infeasibility of implementing a calculation, given that thousands of possible root nodes in certification paths on the Internet exist [47,49, 50]. Each of these PKI's has a different credential enrollment process and different attributes [62]. Thus, a more simplified view can be used to help a service provider calculate its initial belief in the credential presented.

Similarly, Reiter and Stubblebine recognize that neither the computation of paths [2] nor peer trust [62] is sufficient for establishing an initial estimate of confidence in a credential [9]. The researchers conclude that inferring binding of a certificate or credential to a specific entity causes any model to lack explicitness, and a model's meaning should be ambiguous [9]. Following Reiter's lead, this research formalizes an identity belief model by adding factors that make binding explicit, such as enrollment and transmission approach to credential.

Furthermore, the user's opinion of the identity provider is another important principle to consider when describing the belief of an identity and that is the user's opinion [9, 62]. In most research, the user's opinion is a value set by an agent or specific user based on their subjective view of the credential or identity provider. This research sets that opinion as a preset value from the service provider to describe the veracity of the identity provider. Further, Reiter and Stubblebine combine these factors to establish meaningful principles, including simplicity, effectiveness, meaningfulness, and computational feasibility [9]. Evaluation of the belief model presented in this dissertation will draw from these principles and evaluation criteria.

Another criterion often taken into account in belief of an identity is the ability to verify the user's credentials [13, 23, 35]. The model in this research focuses less on belief than the other factors previously established for computing the metric. This research makes the critical assumption that the identity provider has already verified credentials prior to the computation of belief. The next factor often considered in belief models is locality of opinion. Although previous research examines this factor from a more general perspective, [1, 3, 4, 9] Blaze, et al. [23] focus on the need for making a local decision about belief. There are many reasons for this focus, including flexibility of behavior, computational feasibility, and independence. The characteristic of locality in this belief generation approach will allow service providers to maintain independence from external sources, which will be useful in cases where the belief is calculated without external connection. This will also allow quicker computation of the initial belief metric since service providers won't be expected to follow a PKI chain for verifiability.

In addition, Barber and Kim provide a model for establishing a foundational belief in identity [14]. Their agent-based approach has one agent calculating trust in other agents, and representing that calculation as truth and belief. The model components include a view of the agent making the calculation, or a perspective modeler. This modeler represents initial truth or belief about another agent by maintaining a knowledge base, acquiring knowledge about other agents, updating these models, making inferences about the updates, and revising again. The differences between Barber and Kim's approach and that of this research are focused in two areas: data type gathered to form the knowledge base and frequency of update. Barber and Kim's agent gathers the knowledge for the model based only on relations of other agents, while this research will use data to make inferences about the utility of the data to form a belief model. In terms of the frequency of update, Barber and Kim's agent updates constantly, as required by other

agents to keep their knowledge current. The belief store in this research will update at t=0 of a transaction as presented with credentials. This approach is meant to take into account the nature of a user's identity as described in broad terms and not react to the dynamics of transactions.

In order to represent initially and update the knowledge base, this research relies heavily on transactional context. Many Search Engine Optimization and Analytics providers on the Internet have begun collecting this data and storing it for different purposes [66, 67]. This data can also be applied to the present research, as illustrated by a table of commonly used data collected during general web based transactions by Google Analytics [66].

Geolocation
Referring URL
Cookie Presence
Browser Info
Client Environment
Number of Visits
Time Zone
Custom Variables-added by user

Table 2: Google Analytics Transactional Context

Krishnamurthy and other researchers argue that these parts of the transactional context containing personally identifiable information makes them potentially harmful to Internet users [68]. However, this dissertation argues that from a service provider's perspective, this PII can also be used in the computation of confidence in a user's identity.

Trustworthiness using belief-based approaches correlates multiple perspectives to provide a view into the authenticity of a user. In short, calculating the authenticity of the

user can satisfy the quality of belief in architecture with only a singular perspective. Authenticity can leverage the aforementioned transactional context and determine its validity by checking it against the identity provider's golden records for the identity.

2.2.2 Reputation

One of the areas in which reputation has been frequently studied is in agent-based systems. Reputation-based trust modeling is used typically when the trusting agent or algorithm has no experience with the identity it seeks to trust [2]. It is also useful when the agent or user being trusted changes frequently as is the case in systems such the focus is on high atomicity in packet based transactions. Reputation calculation often times includes a method for sharing reputation of users [3]. Another concern in reputation based trust mechanisms is the accuracy of the algorithm. This is dealt with by authentication methods through formal paths and algorithms by digital signatures and chains of trust. The challenge with this approach is that it predicates the reputation of the entire system on the validity of the third party trusted Identity provider. If this changes as the literature on reputation-based trust suggests that it can [9], then it invalidates the entire chain of trust. As discussed above, an agent based or decentralized approach such as in PGP is one possible solution [5]. This research illustrates another possibility to make the reputation of third party identity providers more explicit.

Metrics used to characterize focus on satisfaction of agents, or networks. Behavior-based trust methodologies have also applied Fuzzy Logic [2], which allows the node building trust to establish membership functions that establish a trust model gradually without the need for instant input. Adaptive Trust Negotiation (ATN) is a technique that can allow behavior-based strategies for trust to filter misinformation while querying for new information, as necessary [3, 41]. ATN is extremely useful in this

space due to its ability to help shape and better understand behavior patterns of actors using identities and credentials.

Finally there has been work on defining a category of trust that applies to future interactions using Bayesian Trust Methodologies [37]. Each node will build a graph model of trust of its peer nodes in a directed graph that allows it to conduct analysis to determine which nodes are trustworthy. This is useful in determining context of identity users in order to better establish their behavior in large-scale deployments.

Trustworthiness using reputation-based approaches tends to focus on single agents or peers trusting each other by performing consistent actions. It is easier to generate a reputation for a single identity in an architecture where a service provider is trusting an identity. Calculating the reputation of the user can satisfy the quality of reputation in architecture with only a singular perspective by looking at the repeated actions that the identity takes. The reliability measurement can leverage the aforementioned transactional context and determine its validity by checking it against the identity provider's golden records for the identity. Further the reliability measurement ensures that the same attributes are used correctly over several transactions. If they aren't used consistently and correctly, the reliability metric will be able to detect invalid use.

2.2.3 Combining Modalities

Generally, research shows that combining multiple modalities when making a confidence decision or recommendation is helpful in many domains [3, 13, 26]. In fact, many believe combining modalities is necessary to gain a sufficient picture of the trust associated with any one node or identity. In the combinatorial approach of expert fusion advanced by Verlinde et al., provision is made for static factors [27]. In the case of identity, some known information about the backgrounds of users, static trust

information, or even risk data input into the algorithm to provide better fidelity of information. Verlinde makes the case that expert opinion is a necessary component for any fusion of data for analysis [27]. Others assert that static input must be considered in any automated fusion system [28, 29]. As in other dynamic systems, the types of trust being fused in a multimodal approach cannot account for structural changes in larger system [2].

2.3 Systems in Manufacturing Environments

While the present research is conducted in a transactional context, examining previous studies conducted in manufacturing environments are important for studying identity; these systems functionally represent identities in a fully automated environment instead of users or people. In manufacturing environments, most of the systems employed are categorized as Supervisory Control and Data Acquisition systems (SCADA). These systems have classes composed of systems by function in a server/master and client/sensor/actuator model [78]. Security has often taken a backseat to communication design for performance [76]. But recently, as the Internet has become ubiquitous in manufacturing environments, and deregulation has hit manufacturing industries, security is much more of a concern [76]. Most security issues are driven by the lack of ability to identify systems internal to the network and subsequently authorize them to perform necessary functions [77]. In a closed environment, the identification of systems is relatively easy and can be done through preprogrammed codes and credentials such as certificates [78]. Unfortunately, this strategy is insufficient given the large number of insider threats and rising value of infrastructure run by SCADA systems.

Historically, these insider threats have a few main categories. The first category is changing data values. One of the most important functions performed by SCADA

system is sending or receiving data [76]. For performance reasons, these interactions are almost never protected by traditional means of communicating securely over an untrusted network. Manipulating data values can be disastrous in a manufacturing environment, causing faulty readings in the control center and potentially causing system operators to incorrectly perform functions.

Another type of threat is to change control signals from the master or control areas to sensors, actuators or other types of systems that perform a necessary action. This threat is particularly vexing as demonstrated in the payload of worm Stuxnet which targeted specific models of systems control signals [80]. Stuxnet was able to modify these control signals on the wire by impersonating an authorized system leading to Siemens control systems in a manufacturing environment operating out of tolerance. The next type of threat is focused on fraud by manipulating readings in a manufacturing network for a fraudulent purpose [76]. These insider fraud attacks could focus on either stealing information or detecting patterns that could be used to drive the manufacturer to perform an action.

Attempts to secure these systems have focused in two main areas: the perimeter, and centrally in the environment [76]. Perimeter security is focused on detection of external breaches into the inside system. Perimeter security is solely focused on detecting intrusions from external sources and can effectively detect attacks that are ongoing if patterns can be recognized [80]. However, this security approach is ineffective for the insider threats previously discussed since they all occur inside the perimeter. Central detection is performed by consistency checks of the internal network of systems. These consistency checks will look into the data being sent between systems to ensure that it falls within the modeled expected ranges. Central detection requires that the model of the system be correct and also relies on communication paths to the central

detector being trusted. In most cases within manufacturing and SCADA environments communication paths can't fully be trusted due to the promulgation of IP networks and multiple entrance points to the Internet.

Others have attempted to solve the challenge of securing paths and identification of devices through standard authentication, authorization and cryptographic techniques available on the Internet today [79, 80]. These approaches have some innate limitations due to the manufacturing domain [78]. One domain specific feature is power. Often sensors and other systems operate only on battery and tokens or other credentials drive up power requirements due to frequency of use or other physical characteristics. Another is that for systems to use a password or other credential it must be manually installed due to lack of user operation. This causes the credential to not change for long periods of time. It also precludes the ability to perform password reset, revocation of credential or other techniques that are used to assist in information security. In order to overcome this limitation, some research suggests that adding a third party to trust credentials can help [78, 79].

It is also worth mentioning that another threat vector is a multiple system impersonator performing so called Sybil attacks [79]. These attacks focus on a system Identity impersonator attempting to authenticate with multiple identities until it finds one that works. Sybil attack detection has been proven possible with enough knowledge and a lot of resources available [80, 81]. Further, Douceur postulates that Sybil attacks are never fully preventable without a trusted third party [81]. The question remains, if it is impossible to trust a third party to verify credentials then is a Sybil attack always possible? What are the ways to reduce its possibility of occurrence?

2.4 Value of Identity Attributes

An online identity has value to the person using it, the identity provider, and the service providers that use it for transactions. Therefore, the attributes that make up the identity must also have value. First, there is value to the individual for a variety of different reasons. These tend to be related specifically to the attributes themselves. For example, users have a preference for using the same user account across multiple identity providers [86]. The value is derived from convenience as well as personal notions of consistency across communities of interest. Mueller also discusses the costs in time and memory burden of maintaining multiple passwords [86]. Finally, there are switching costs for user account names that involve telling others new information about yourself. Mueller further discusses similar preferences that users have for other attributes by performing a conjoint study on cognitive preferences for different types of identity providers [86]. The result of this study was clear that certain groups of people placed statistically significant values on their attributes. These can be measured based on group type and demographic.

Identity providers also place value on the user identities. There is a robust market for buying and selling identities [48] with some business earning several dollars for personal attributes. Often these sales are harmless rather then contributing to the overall rate of identity theft. Identity providers also realize this and have chosen to select common sets of attributes as documented in the 2014 NASPO report [88]. This report indicates that identity providers tend to choose standard Name, Location, Time and Identifying attributes to provide a very high identity-resolution rate. Identity resolution is important when enrolling users as it verifies that the person establishing the account is who he says he his. Yet, this can become a double-edged sword though during verification. If an attribute is compromised and the same attribute is shared across many

providers, that attribute compromise can have a high impact. The impact will be felt among all service providers who use that particular identity. According to the Verizon annual identity report, compromise of identities and their associated attributes costs service providers over 400M [48]. According to Horton, humans strongly correlate compromise with utility. The more compromised an attribute is the less useful it is [89].

Thus it is critical to study the probability of compromise of identity attributes in order to reason about the rest of the components of Identity. If identities are composed of attributes any abstraction of these attributes will provide at best incomplete information about that Identity. At worst it will provide false conclusions about the validity of the identity or the process used to obtain the result.

2.5 Risk Modeling Methodologies

The use of a methodology to dynamically calculate a risk model based on (at a minimum) the above events would support identity verification and authentication services for calculating confidence. The most recent *Identity Fraud* report indicates that risk isn't used enough in calculation for securing data [32]. Manchala established that much of the information that goes into calculation of trust in entities could also be used to help inform risk models for access to that data [2]. In its most basic form risk is characterized as the probability of an occurrence of something along with its corresponding consequences if it happens [19]. In order to apply risk to the problem of identification, it is necessary to find a generalized approach to characterizing its current and future states. Much of the data that could be collected during an identity transaction could help feed a probability calculation for specific acts. CORAS has demonstrated the ability in specific instances of a model to return risk results for its underlying data [36].

This approach is difficult and very web service specific, as it requires a detailed model including attributes and types of information stored to return information risk.

A more generalized risk in data access described by Kearny, et al. suggests it is important to consider information holistically, and that usage of external risk information should be compared to a generalized case of compromise [19]. To gain a clearer picture of the future state, this research focuses more on a Bayesian approach as suggested by Bonafede, et al. [38]. They suggest using an enterprise wide approach toward assessing a future state of risk. If rebuilt using a belief estimate and reputation as inputs, this model will apply to the enterprise of data secured by identity-based authentication and authorization mechanisms.

Existing risk modeling methodologies fall short on their ability to model risk without understanding the consequence of the event. In the case of Identity, consequences are almost innumerable. Thus it is important to be able to quantify probabilistically a risk that an identity isn't valid. The validity of an identity is sufficient information for a service provider with which to make a decision.

Chapter 3: Identity Attribute Usage

RQ1: Which attributes of Identity are most trusted?

In order to calculate trust in an identity, the attributes that compose it must be considered. It is possible to model these attributes in such a way that allows a provider to reason about which attributes to use and how to weigh them in any sort of usage. The weighting approach will occur in the model presented in this research to enhance attribute utility by minimizing Probability of Attribute Compromise (P_C) and maximizing the use of populated, known attributes. By considering P_C as a separate metric, it is possible for Identity Providers to make dynamic use of P_C to conduct calculation of confidence in Identities that are attem$_p$ting to gain access to resources. In effect, the compromise probability also allows the Identity Providers to prefer attributes. An attribute that had a low P_C value would be preferred over one with a high P_C value. This preference allows the identity provider to buffer itself against possible attacks that could happen based on attributes that are compromised for the same identity in another provider.

When a person enrolls with multiple identity providers on the Internet, each provider may ask for a different set of information or attributes upon enrollment. These attribute types are often duplicated across sites and repositories. These attributes form the basis for identification of a user to service providers. If those attributes have similar values, then a compromise in one identity provider can lead to a compromise in another. Thus, this chapter presents a new model for determining attribute preference based on attribute usage and the inherent connections (dependencies) between attributes. The model is empirically evaluated with two data sets – the Identity Ecosystem and the US Army central identity repository.

Identity providers use a variety of methods to combat is problem of attribute compromise. One method is verifying sets of attributes to ensure the user who enrolls is actually who he or she claims to be [88]. This method has the advantage of providing a strong confidence on initial enrollment and a basis for future comparison. However, once these attributes change, they can become less useful in for verifying an identity. Another method conducting attribute validation at the time of use, often called online verification. For example, Google's runtime position verification for transactions [90]. Unfortunately, this method is a narrow implementation of attribute verification; it only works for specific attributes, such as location. In order to be more useful, this method must be more broadly applicable across attributes.

Understanding which attributes are most likely to be compromised and establishing approaches to mitigate theses compromise is necessary. For a single user with multiple attributes related to each other such that $I=\{A_1,A_2, \ldots A_n\}$, each identity implicitly has the full set of attributes. For example, a user can have a Social Security Number and driver's license. Some attributes are more likely to change than others, such as address. When enrolled, these attributes become instantiated into a credential provider. The resulting set is most likely a subset of the overall attribute set for an identity. As a person registers for multiple credentials with multiple providers with different needs, these sets can become disjointed.

For example, provider N chooses the set of attributes of type $\{A_0,A_5,A_8\}$ to represent the person with identity I_0 whereas provider M chooses the set of attributes of type $\{A_0,A_6,A_9\}$ to represent the same I_0. The attribute of type A_0 is the only attribute common to both of these sets. Therefore, the security functions being performed by provider M using A_0 are potentially disrupted if A_0 is compromised in provider N. This disruption results in the degradation of assurance associated with A_0. Furthermore, A_0 is

connected to the rest of the other attribute types. Due to the connections between attributes and the dependency of credentials on respective attributes, it is possible to obtain other attribute types from A_0. For example, if A_0 were a Social Security Number, then it would be possible to obtain user credentials (full name, citizenship status, etc.) from the compromise of that single attribute. In practical terms, this puts all user transactions at risk.

3.1 IDENTITY ATTRIBUTE MODEL

In order to calculate popularity among attributes, this research models the linkages between them. This will allow reasoning about which attributes are the most likely to be used in an attack. The reference model will follow the same pattern as the Identity Ecosystem, which shows measurable relations between these attributes [18]. The model will reflect an overall identity view of a person and is applicable to different types of identities. Furthermore, the model can be tailored to individual identity providers based on their specific attribute types and connections.

In order to reason about the connectedness of these attributes, Identities can modeled in a graph reflecting attackers patterns as they compromise attributes (i.e. which attributes are used to gain access to other attributes). The graph defining an Identity is represented as a social graph [100]. In this dissertation, the nodes are represented as identity attributes and the directional relationships between nodes reflect that one attribute can be discovered given the other. Thus, the relationships describe that an attacker can compromise can one attribute given a certain known attribute. For example, if the attacker knows a person's name, there is a probability the attacker can discover the person's address. The relationships in the graph are weighted by these probabilities of discovery/compromise. Let the nodes of the graph represent attribute $I[A_n]$. It is

asserted that the node's value is independent of other nodes around it and it can take on values independently. This makes sense in the context of a person's identity because an attribute must be fully independent of another (as defined in Chapter One). A person's name value is independent of their address. Let the relationships of those identity attributes be represented by edges. Each edge in the graph represents a probability that one node will be compromised given the other node. The corresponding weight of the edge type can be strong or weak. For example, it is possible to obtain a person's full name from a Social Security Card; therefore, the likelihood of obtaining one from the other is high. The Center for Identity Ecosystem [18] characterizes different types of relationships as shown in Table 3. In some cases, attributes have no relation; therefore, attributes in this model are shown as unconnected. This research will leverage the "Probabilistically_Determined_By" weighted relationship meaning one attribute, A_1, points to another attribute, A_2, with weighted probability, say .6, if attribute A_2 can be discovered with 60% probability given attribute A_1.

Type
Change_Sensitive_To
Temporally_Precedes
Probabilistically_Determined_By
Determined_by
Necessary_for
Accessible_via
Bred_from
Composed_of

Table 3: Properties of Attribute Relationships

Figure 2 (a sample identity relationship graph) provides a better view of the relationships between attributes including the edges labeled with the appropriate transition probability

from node to node. For example node 6 has a certain probability that it can be obtained from node 1 and that is shown as edge P_{16}

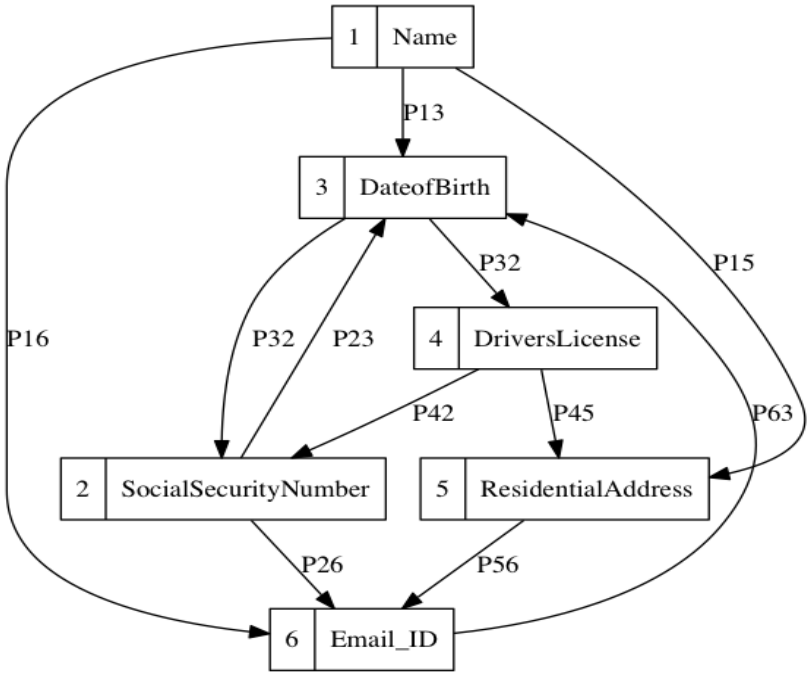

Figure 2: Sample of Attribute/Relation Graph

By representing an identity as a collection of attributes as shown in Figure 2, one represents an identity abstractly. This abstraction focuses on the connections between the attributes for a generic identity. It is possible to use this model to overcome the limitations of looking at identity strictly from a single identity provider's perspective since multiple attributes can be used by different identity providers [88]. The collected attribute types for each instance of an identity are distinct. The edges between the attributes are the only means of showing overlap or relation between each of these attributes. To extend the model to support additional attributes and connections, the process illustrated in Table 4 is leveraged.

Step	Description
1	Consider a possible identity attribute
2	Determine if it exists in the existing model
2a	Check for a misspelling of attribute type and possible names for the type
2b	Check for an attribute that has a dual meaning and update possible names for the type
3	Add a new attribute
4	Model the relations that the new attribute has with existing attributes
4a	Select Related Attributes
4b	Select Edge Type
4c	Add new Edge
5	Repeat until complete

Table 4: Modeling Approach for Addition of New Attributes

3.2.1 Probability of Compromise

Establishing the Probability of Compromise property, or P_C, is built upon a graph for reasoning about preference between identity attributes. In the model of a social digraph with attributes connected to each other, "attributes most likely to be compromised" are defined by popularity. Popularity is represented through a graph centrality measure called Eignenvector Centrality. The "attributes most likely to be compromised" are the most "popular" based on how connected an attribute, say Attribute A_2 is to other attributes and the weights of those connections described by the probability that Attribute A_2 can be compromised given an attribute it is connected to. Thus, more connectivity (popularity) and higher probability weight means the attribute A_2 can be reached by transitioning from many different other attributes with greater likelihood.

Put differently, the more reachable the attribute is from other attributes, the more likely it will be compromised. Compromise is modeled as the ease of obtaining an attribute given the compromise of another attribute. The aforementioned random walk across all of the attributes inside this graph will be represented as a digraph. If an attacker obtains one attribute randomly, he can move randomly throughout the graph compromising other connected attributes without regard to the attribute he started on. The identity graph is also posited to be a closed system. Thus, an attacker looking to compromise a person's identity will attempt to gain more information about that person. The attacker focuses on obtaining one attribute, which can sometimes involve several traversals through the graph.

Another necessary property to consider is connectedness. After assessing graph connectedness in this model, there is one Closed Communicating Class (CCC) with a few attributes not connected to that CCC. Therefore, the node popularity within the graph can demonstrate steady state conditions. The analysis on the steady-state condition of the digraph ultimately provides equivalence to Probability of Compromise for each attribute. For the purposes of this digraph, it is only necessary to understand what transition probabilities between the attributes exist. The transition probability between each attribute is based on the edge transition probability, represented by the weight of the edge. A total probability of transition from node i to node j is shown in the below equation where $I(P_{ij})$ represents presence of an edge between nodes.

$$I(P_{ij}) \begin{cases} 1 \; if \; P_{ij} \; has \; an \; edge \\ 0 \; if \; P_{ij} \; has \; no \; edge \end{cases}$$

These edges are directional. This formula will be called the transition probability generator or TPG. The Transition probability between nodes *i* and *j* is then represented by the below equation.

$$P_{i-j} = I(P_{ij}) * P_{ij}$$

In order to derive P_{ij}, one must understand the probability of compromising node j from node i. In order to come up with a more concrete prediction, existing data was used from The Center for Identity Threat Assessment and Prediction (ITAP) model. The ITAP offers data to populate the social graph to define an identity. The ITAP represents the processes by which criminals steal and fraudulently use identity attributes. Currently, there are about 2700 stories captured in the model. This research automatically extracts attributes from the ITAP and represents those attributes as nodes in the social graph. The edges are created by an exhaustive search of the ITAP theft and fraud stories describing the process steps by which identity attributes are used to compromise other attributes. The probabilities along the edges are calculated as the likelihood that attribute j is compromised given attribute i. In other words, the weighted edge from attribute i to attribute j is calculated as the number of times that a certain attribute j, is obtained/compromised given another attribute i divided by the total number of times that attribute i is used to compromise other attributes including j. These probabilities are represented as the weighted values on each edge of the graph above (Figure 2). The equation that supports P_{ij} is listed below:

$$P_{ij} = \frac{Occurences\ of\ j\ obtained\ from\ i}{Total\ number\ of\ occurences\ of\ i\ as\ a\ origin\ of\ compromise}$$

All of the edge probabilities exiting each node sum to 1. Thus, the digraph model for calculation of node centrality fits the eigenvector centrality measure. This technique is applied to the social graph to show node preference, which this research defines as the attribute node most likely to be compromised.

As previously discussed, these transition probabilities can be related in a transition matrix within the bounds of the digraph. This transition matrix combines all of the values of transition probabilities from the TPG's a single matrix shown below. This matrix becomes a square matrix because each node in the graph is related using the TPG to every other node in the graph exactly once. Note that the due to the attribute model and the derivation of P_{ij}, each diagonal term is set to 0 since it is impossible to remain in a state.

$$P_{ij} = \begin{bmatrix} P_{0-0} & \cdots & P_{0-j} \\ \vdots & \ddots & \vdots \\ P_{i-0} & \cdots & P_{i-j} \end{bmatrix}$$

Once modeled, the matrix is fully populated with the transition probabilities between all nodes. Since all probabilities from each i must sum to 1 the process can be considered stochastic. Once the matrix is determined to be stochastic it can then be used for popularity computations. Some further analysis is required to determine its suitability for computation. After analyzing the transition matrix P_{ij} the period is found to be 1; thus, the model is aperiodic. Next, the model was analyzed for recurrence. It was determined to be positively recurrent because there is a finite probability of returning to each state in the matrix. Thus, the transition matrix P_{ij} is found to be irreducible, aperiodic, and positively recurrent. This means that the matrix is not transient but converges at some time before infinity.

In order to express the influence of each node in the network, the adjacency matrix Pij is written as the eigenvector equation below:

$$P_{ij}x = \lambda x$$

The Eigenvector of the matrix P_{ij} represented by λ is the steady state probability of landing on a node in the graph. The vector λ represents the state space of the solution to the equation above.

$$\lambda = \begin{bmatrix} P_0 \\ \vdots \\ P_i \end{bmatrix}$$

Since the P_C is solely dependent on the connections between the attributes, it is important to ensure these are modeled correctly and that the model is flexible enough to accommodate their change. The identity provider would like to optimize these attributes such that any single attribute that is commonly collected during the enrollment process across multiple identity stores is not weighted too heavily in authorization decisions. For example, an email address is more likely compromised than most other attributes. This also means there is a ready path to most of the other attributes.

3.2.2 Probability of Fill

The Probability of Fill property, P_F, is a means to describe the usage of attribute types. The below equation represents the probability that a single attribute value will be filled. The variable m_i represents the attribute type and is set to 1 if a particular instantiation of the attribute type m is set. The identity repository has n instantiations of attribute type m.

$$P_F = \frac{\sum_{i=o}^{n} m_i}{n}$$

Therefore the probability of fill is a simple weighted average of all of the customers in an identity repository that have the attribute.

Liu expresses the importance of understanding how often a single attribute type is used in the computation of trust [6]. While Liu focuses on direct interactions of agents, the general concept is applicable to the model in this research. The property can be calculated as a single or across multiple models, meaning the property is useful for one or more credential providers. If a single credential provider doesn't use an attribute, then others might choose to and the lower P_F drives the preference lower for that attribute due to less usage. It adds information related to usage that is not used in the P_C metric. P_F is a representation of the number of attribute types in use in a single model or single identity repository. For example, if a Social Security number is used for every account inside the model, then the SSN attribute type node has a P_F value of 1.

Probability of fill can also apply across identity repositories. If a single identity is considered, the same equation can be used with *n* representing the number of repositories the identity is stored in. Without considering the fill of attributes, compromise could not be effective in the computation of attribute preference. If compromise were only used to calculate preference, then it would be possible to prefer that is low risk for compromise but is never used. This could provide a false motivation to an identity provider to prefer certain attributes

3.3 EXPERIMENTS

The following experiments were conducted to determine the effectiveness of the P_C and P_F metrics in providing attribute preference. The combination of these metrics will allow identity providers to more properly weight attributes based on their fill and

compromise characteristics. This weighting will allow dynamic decisions about which attributes to trust. The Identity Ecosystem dataset was compared with a dataset that is derived from a US Army identity repository [101]. The high level details of each dataset after modeling are shown in Table 6 below.

Metric	**Identity Ecosystem**	**Army Attributes**
Number of total attributes	104	18
Number of edges	1435	104
Highest Probability of Compromise	.439	.2348
Expected Value of Distribution	.0445	.02326
Attribute with highest P_C	Social Security Number	OrganizationalUnitName

Table 6: Difference in Attribute Models

The US Army identity dataset contains a small number of attributes. In order to model and analyze it for this research, experts were interviewed to determine the edges or connections between the attributes [102]. This dataset was chosen for two reasons. First the number of attributes was small. The small number of attributes can help to determine the similarities of a small-scale model to the ID Ecosystem. Secondly, the attribute types were of different syntax than the ID Ecosystem model. This overlap provides an opportunity to update the ID Ecosystem model with new attributes to add to its richness..

As indicated in the modeling section, in order to be useful a model must be able to be continually updated.

3.3.1 Experiment 1: Calibrate P_C

The modeling process in Table 4 was followed to perform the probability of compromise calculations for the two different attribute sets –attribute set from ITAP and attribute set from U.S. Army. The first of these sets is based on the ITAP data [103]. The ITAP data looks across over 3000 compromises of Identity Attributes. The compromise probabilities were based on the formulas in section 3.2.1. The result is a calculated Probability for Compromise for each attribute based on the available data in the social graph created from the ITAP data. Based on the modeling activity, some attributes are not connected and thus factored out of the analysis, leaving a closed communication class on the digraph. Figure 3 shows steady state compromise probabilities for each attribute. The attribute number identifier is on the x-axis and the steady state compromise probability on the Y-axis.

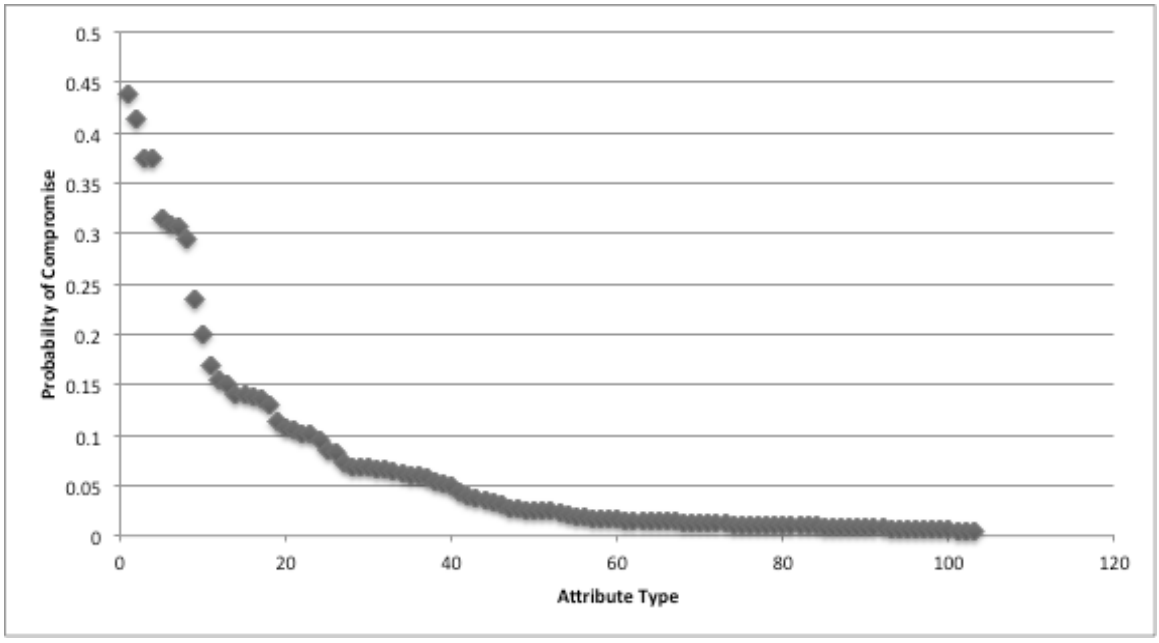

Figure 3: Attribute Preference for Compromise

The calibration of the Identity Ecosystem shows that there are a small number of attributes with a higher probability for compromise. These attributes are especially prone to be compromised as shown in Table 9. An Identity Provider could choose to change its process of identification of its users to focus on those attributes, which are at lower risk for compromise. This is important because it can help to stem the significant desire of Identity providers to continue to collect more and more information. Also, eventually the more an attribute is used, the more likely it is to become compromised. This model can allow an Identity provider to reason about emerging attribute connections and change preference to match the attributes with the lower P_C values.

In a similar fashion, the US Army dataset was considered. The US Army dataset reflects attribute information taken directly from the Army Knowledge Online repository, which currently functions as the authoritative data source for the U.S. Army. The compromise probabilities were based on the formulas in section 3.2.1. The result is a calculated Probability for Compromise for each attribute based on the available data in the social graph created from the Army data. Based on the modeling activity, some attributes are not connected and thus factored out of the analysis, leaving a closed communication class on the digraph. Figure 3 shows steady state compromise probabilities for each attribute. The attribute number identifier is on the x-axis and the steady state compromise probability on the Y-axis.

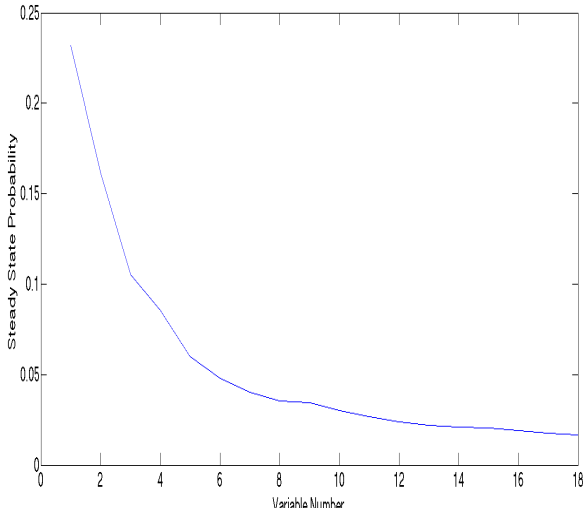

Figure 4: Trendline Steady State Compromise Probability Army Dataset

The calculations using of the US Army dataset similarly shows that there are a small number of attributes with a higher probability for compromise. These attributes are especially prone to be compromised as shown in Table 9. With this information, the U.S. Army could consider changing its process of user identification to focus on those attributes, which are at lower risk for compromise. This is important due to the cost of collecting and updating attribute information as well as the consequences of stolen (i.e. compromised) identity attributes. The fewer attributes collected and used, the more money that can be saved by the U.S. Army. If those attributes have a lower Probability of Compromise, the U.S. Army also incurs less risk with those collected attributes. Also, eventually the more an attribute is used, the more likely it is to become compromised. This model can allow the Army Knowledge Online to reason about emerging attribute connections and change preference to match the attributes with the lower P_C values.

3.3.2 Experiment 2: Vary the Number of Edges connected to a Node

In order to demonstrate that the model provides a Probability of Compromise that varies as connections to attributes vary, the below sets of experiments were conducted. These experiments verify that the model is sensitive to the changes in identity attribute relations.

Experiment 2a- If the number of edges with an indegree to any node is decreased, a corresponding change in the P_C value for the attribute will occur.

Conditions- The experiment is run first on the highest connected attribute decreasing the number of edges it has to other nodes.

Results- As the number of edges to an attribute decreases, the Probability of Compromise decreases as shown in figure 5. Further it is noted that the highest P_C value in the graph changes between attributes and varies independently of the decrease in node P_C value.

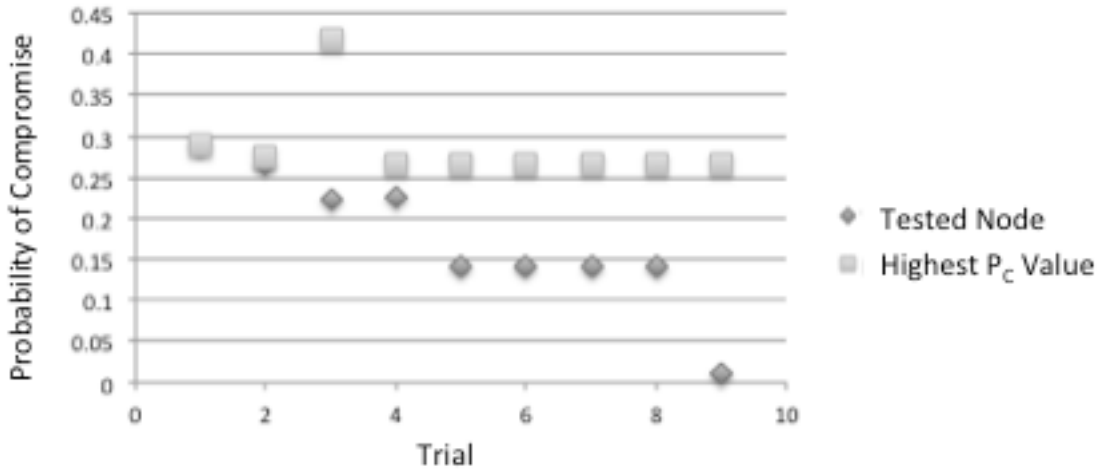

Figure 5: Probability of Compromise for InDegree across 10 trials

The results show that as an attribute is less connected to other attributes, the Probability of Compromise goes down. In order words the less paths that can be taken from the attribute to a random attribute will lower the Probability of Compromise. The decrease in Probability of Compromise verifies the model expectation that as fewer attributes have an ability to derive the value of a connected attribute there should be less compromise.

***Experiment** 2b*- If the number of edges with an InDegree to any node is increased, a corresponding change in the P_C value for the attribute will occur.

Conditions- The experiment is run first on the highest connected node on each attribute increasing the number of edges going out to other nodes.

Results- As the number of edges going out of an attribute increases, the Probability of Compromise for that attribute increases as shown in figure 6. Further it is noted that the

highest P_C value in the graph changes between attributes and varies independently of the decrease in node P_C value.

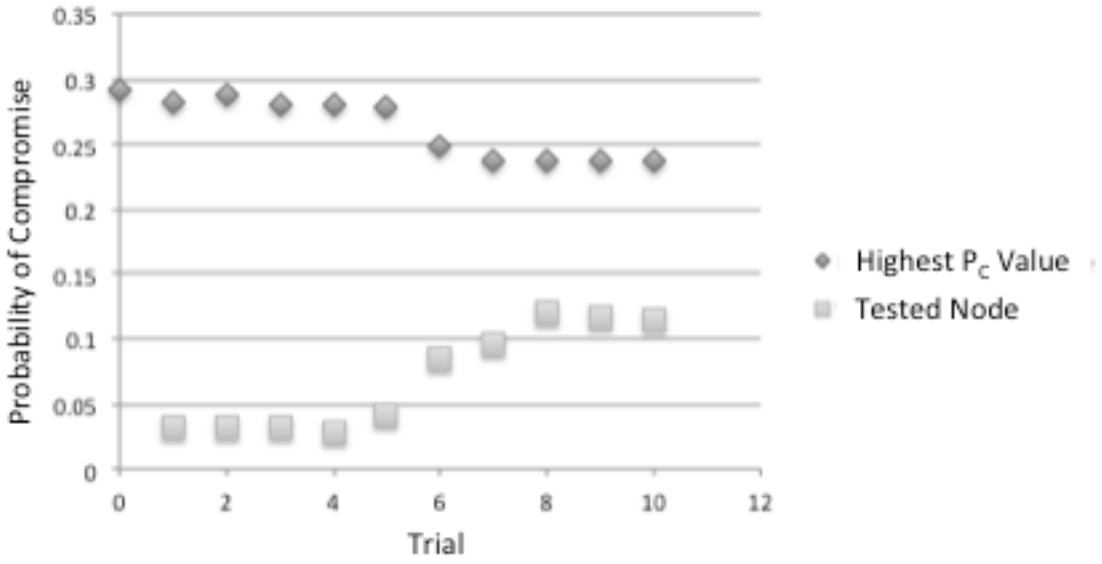

Figure 6: Probability of Compromise for InDegree Across 10 Trials

The results show the converse of the first experiment is true. These results indicate that an attribute that is more connected has a higher Probability of Compromise. The increase in Probability of Compromise verifies the model expectation that as fewer attributes have an ability to derive the value of a connected attribute there should be less compromise.

The behaviors of attributes in the graph demonstrate some useful properties. The first property is that the highest probability attribute in the graph doesn't fluctuate significantly when edges are changed. This can be useful for identity providers to determine their preference for attributes. If they are able to adjust their approach to focus on using attributes with lower probabilities for compromise, they can be assured that the attribute most at risk for compromise will not fluctuate significantly. The lack of top end fluctuation cannot be guaranteed with the empirical method used. Since, the real world

attributes have a significant risk of compromise, with enough work it would be possible to raise the compromise probability of a specific attribute very high by a significant degree of compromise in a measured time period. The change in probability wouldn't raise the relative ranking of the attribute necessarily but it makes it less computationally feasible for Identity providers to use in the choice of attributes for authentication.

3.3.3 CALIBRATE P_F

Since P_F is used across multiple identity repositories, it becomes a useful detector for demonstrating how much an attribute is used. For example, if identity repositories 1 and 2 require a firstname as part of the enrollment process, then one would expect that P_F for that attribute would approach 1 given that it is required. As P_F scales to multiple identity repositories the metric appears to have a linear relationship. Figure 7 shows the relationships of P_F across 4 different hypothetical identity providers.

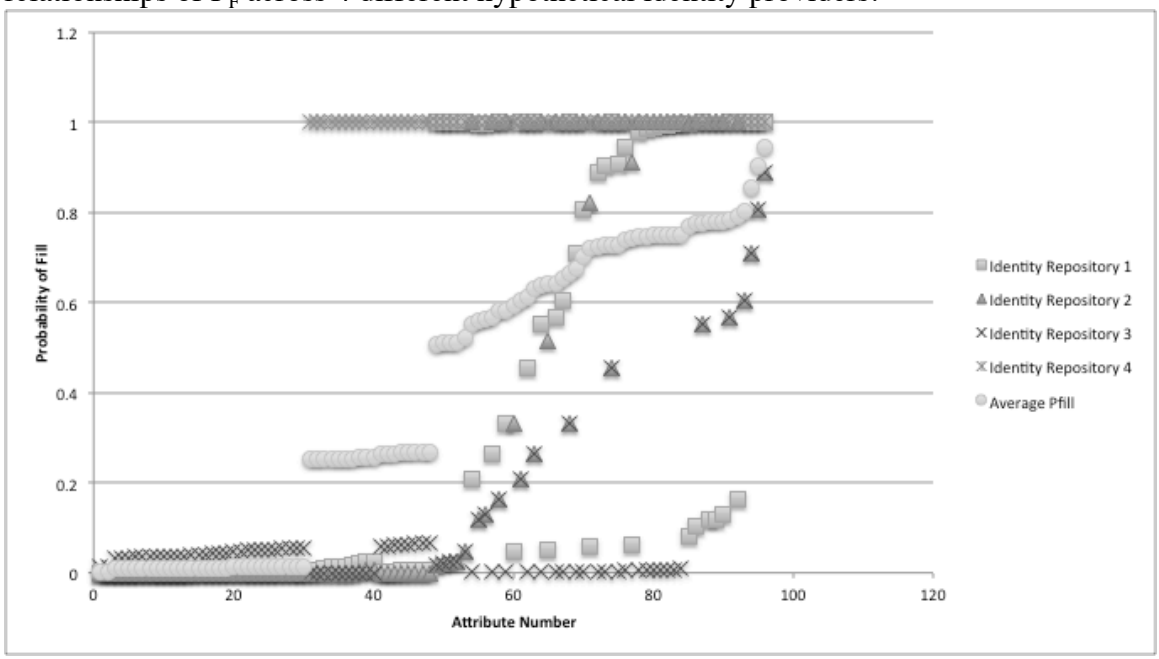

Figure 7: Relationships of Attributes with Differing P_F values Across Identity Providers

Many of the attributes in use are not likely be filled, which indicates that either identity repositories do not present the ability to fill these to users, or they are unpopular. The use of P_F is local and only effective when a large number of providers report their statistics.

The P_F value for an attribute is somewhat difficult to calculate in practice due to the need to understand the run time usage of these attributes within an identity provider. In theory however, fill probability is very useful. As demonstrated in the experiment, it helps to inform identity providers about which attributes are most used. This is a useful property when combined with the attributes that are most compromised. If for example, an attribute provider were to prefer an attribute with a lower probability for compromise, but high fill probability, then there is a higher chance that if that attribute were to be compromised anywhere it could be used to exploit the identity then if there was an attribute with the same compromise probability and a lower fill probability. Both metrics are useful for calculating attribute preference. In the research reviewed, probability of fill was not taken into account in determining the preference of an attribute. For example Blaze et al[23] is concerned about the usage of attributes by identity providers with respect to preference but conveniently leaves out the idea of measuring it. It appears that most current research doesn't use this metric because it is hard to define practically. It is however possible for Identity providers to build a means of sharing this information with each other without revealing specific identities, trade secrets or compromising details about their inner workings. That is a possible avenue for future work.

Chapter 4: Identity Trust

Attribute preference alone is not sufficient to answer the question of whether an identity is truly trusted. One must consider the question of usage of the identities that contain the attributes. Without any sort of approach for calculating authenticity of an identity, the only source of trustworthiness in the assertion is whether attribute types match in value. If an attribute can be verified such as SSN then the identity is deemed trustworthy. Leveraging existing information about the identity and the attributes used to enroll that identity helps to add a degree of fidelity to the trust calculation.

RQ2: Can the reliability and authenticity of an identity in a transaction help demonstrate trust to an identity?

Building on the RQ1's calculation of attribute preference, RQ2 leverages available information at the start of a transaction. With a better idea of which credential attributes are preferred, the calculation of trust can compute authenticity with a comparison to specific attribute considerations such as fill and compromise. This comparison yields a factor representing the trustworthiness of a credential's source based on the types of attributes that it uses. Another comparison can be made between the values of each of the attributes. If the values of attributes (e.g. for an instance of an SSN, person's name on SSN card) in one credential supporting an identity don't match the golden records within the identity provider, the overall trust probability will be smaller. This chapter explores these and other comparisons.

4.1 Example Scenario for Trust Calculation

Based on the Reiter's principle of trust calculation [9], we assume that a user may gain access to resources by presenting a credential, but acknowledge that a user does not necessarily own the credential. To illustrate this concept, we use the hypothetical

example scenario of "Alice," a consultant with a United States federal agency (USFA), specializing in software analysis. USFA employees have access to different resources in a cloud delivered by ABC Industries, a security-conscious private company. Alice and her collaborator Bob both have credentials from ABC Industries. While working on an important project together, Alice shares her credentials with Bob and asks him to retrieve critical files from the cloud. Now, regardless of his intent, Bob has compromised Alice's identity.

And let's say that Alice and Bob travel to a customer site and work together from a hotel. Mavis is in the room next door and uses simple, publicly available tools to capture hotel network traffic and intercept Alice's credentials to ABC Industries. Mavis uses the intercepted credentials to access ABC Industries' computers and resources while impersonating Alice. Further, Mavis is able to use Alice's credentials to set up other accounts and impersonate Alice on social networking sites. Other identity providers may use these accounts to verify Alice's identity, and so on.

Overall, given the critical nature of the resources stored in their cloud, ABC Industries needs to improve existing identity verification methods to ensure that Alice and Bob are who they say they are. Such improvements will aid ABC Industries in granting Alice and Bob the access they need to perform their separate functions, while still maintaining the integrity of their individual user information. Such a scenario seems to be commonplace in online transactions, and multiple solutions have been developed to increase user-identity trustworthiness based on credentials, including encryption, multiple factor tokens, and secured networks. However, these solutions fail to solve the fundamental problem of authenticating a user's identity: How can we ensure that a user is who he or she says they are? Therefore, this research makes the following assumptions

in creating a new method for calculating user-identity trustworthiness from a provider's perspective:

- Trust is bound to a credential. If any attributes change, trust in the credential must also change.
- Trust in the user is transitive throughout the system; trust in a user is passed on to the resource via the identity provider.

Based on these assumptions, this research posits that users can be identified with greater confidence by calculating trust using transactional context and behavior.

4.2 IDENTITY TRUST CALCULATION

The following section discusses an approach for calculating user-identity trustworthiness. The attributes identity providers collect during an online transaction are compared with reference data about the offered identity (i.e., the identity a user offers during enrollment). These comparisons yield a probability representing the trustworthiness of a user's identity based on the frequency, types, and values of attributes.

4.2.1 Model-Based Approach for Trust Calculation

This section provides a model that can be used as a frame of reference for evaluating different components of user-identity trustworthiness, each component is represented as a graph. Using a graph-based approach allows the model to interconnect seemingly unrelated concepts and perform computations. For example, if attribute type A_1 has a certain value n, then that value can be mapped to different types of attributes, transactional contexts, and identity indices. These connections have properties that will allow for the direct computation of authenticity and reliability.

Transaction. The first component to model is the transaction. Let U_1 be the node (person or entity in a network) initiating the transaction, and U_1 communicates a transactional context $[A_0 \ldots A_n]$. Let I0 be the user identity. Let the identity provider maintain a golden record with all of the attribute types and values used when U_1 enrolled with the identity provider. The transactional context is composed of the user's identity and its associated attributes (see Figure 9). It is important that the transactional context is indexed by an offered a declared identity such that it can be used in subsequent calculations to ensure consistency or detect errors.

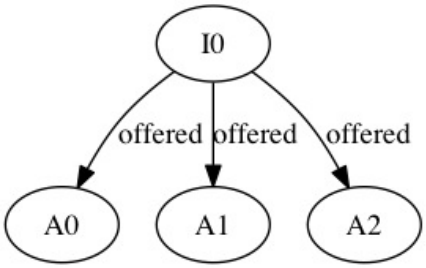

Figure 9: Transactional Context

The identity provider maintains data about the relationship between values of attribute type A, and the associated identity, I0. Different levels of user-identity trustworthiness can be calculated by analyzing these relationships (see Figure 10).

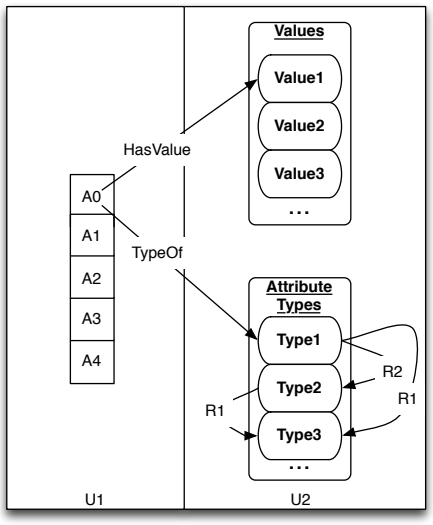

Figure 10: Transactional Context and Reference Attributes

In this case, the trust calculation will be focused on unidirectional trust from the identity provider to U1. U1 can present some or all of its attributes to the identity provider. The identity provider maintains a model of indexed identities, user attributes, and values. Indexed identities are simply connected attributes, and each attribute has a type and value as reflected in the below sample. The notation A_n is an attribute type and each bracket shows values. These nodes are strongly connected through the relationship property ISCONN, which is bidirectional but represented as two directional edges as shown below:

A(59){name:Essie}-[r:ISCONN]-A(207){name:Vaill}-[r:ISCONN]-A(33){name:Anchorage} - …

Each identity stored within an identity repository is represented with a set of nodes and relationships similar to the above notation. Each attribute type A(x) is represented by a

node, and connected by edges with all of the different nodes containing values for that type. Thus, it is easy to compute the number of times an attribute is used by summing the edges that have the relationship *TYPE_OF*.

Attribute. The next identity aspect to model is the attribute. Each attribute type is represented as a node in a graph. These nodes are connected with relationships representing the interdependent nature of identity attribute types. A portion of the graph is shown in Figure 11.

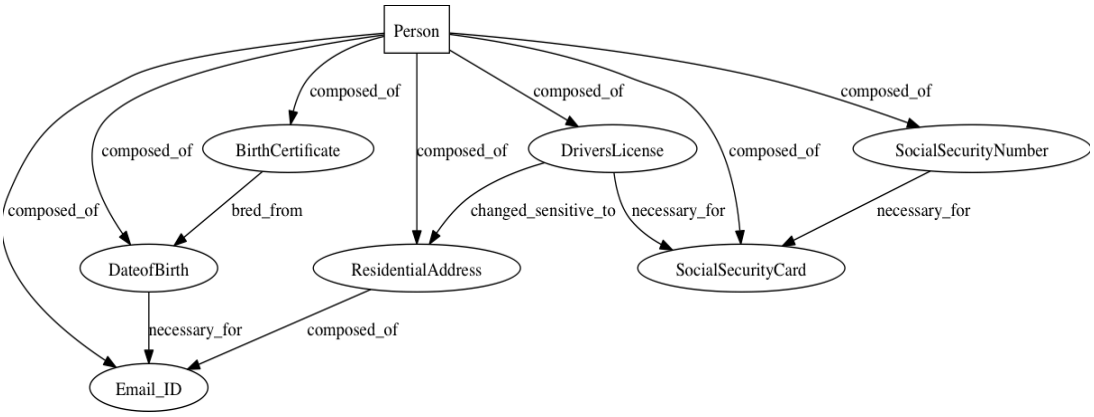

Figure 11: Identity Attribute Relationships

The graph contains a root node, which provides an index point of all of the other nodes. These nodes each represent a different identity attribute. Each attribute node is one edge of the root of the graph. Each attribute has some type of relationship to another attribute, and these relationships are weighted. For example, a Social Security number is necessary to receive a Social Security card. The value of the *necessary_for* relationship can be modified as necessary to fit the particular type of computation. When analyzed, the attributes, and their associated edges and relationships, can be represented as a probability matrix [95]. This matrix offers the ability to answer questions such as:

- Which attribute is used most frequently?

- Are there certain attributes that should be protected more than others?

4.2.2 Trust Algorithm

To successfully decompose the available transaction data into measures and subsequently allow computation, the following algorithm is proposed:

Algorithm 1. Trust Generation Model
Input: Node Set $\{A_0 \ldots A_n\}$ called TC_1
Output: The Rel Ai, Auth Ai.
index = 0; Num_α =n;
repeat
TC_α = Findsubset(TC_1, *index*);
Found = *TRUE*;
for *each node* A_i **do**
ValA_i= *Count Same* A_i *Values*);
NumA_i=Count all Values of A
AuthA_i=ValAi/NumAi
Return AuthA_i
for *each node* A_i **do**
ValA_i= *Count Same* A_i *Values*);
NumA_i=Count all Values of A
RelA_i=BinomAi(ValAi,NumAi)
Return RelA_i

until $FreNum_\alpha$ > n;

A transactional context is available at the beginning of a transaction and conducted with a set of attributes $\{A_0 \ldots A_n\}$. These attributes have values and are categorized by types such as Social Security Number, Mothers Maiden name, etc... during pre-processing. After transactional context is obtained, the list of the golden record for the identity is obtained by searching the graph using the People index at a depth of 1 to all of the reference ID nodes. The search is represented as:

For all Identities, In, Return number that contains types A_n with matching values

When an identity is read, its value of access time is updated. If no matching identity nodes are found, then the algorithm creates a new identity reference node and adds the associated attribute types and values. These attributes are linked to the reference for their type and value. Each attribute type is checked for value matches inside the set of reference identities. This number of matches is returned and used in the computation of authenticity. For practical reasons in a graph model, a set corresponds to a set of connected nodes with edges that take the value ISCONN. Thus, during a depth-first search of the graph, each of the TC attributes are compared to the reference type, traced to reference index nodes, and confirmed as connected.

Finally, once Algorithm 1 has run, computations can be made to determine a belief metric for the identity attempting the transaction. Neuman et al. [93] uses a weighted Dempster-Shafer approach combining reliability, availability, and promptness. These authors posit the approach can be extended to combine any two human characteristics of trust, as long as each has a probabilistic value and the approach specifies the uncertainty of the characteristics. Let the universal set be Θ. Reliability and authenticity assign probability values over this universal set. The below treatment applies the Dempster- Shafer approach to the computed values for authenticity and reliability. The following equations derived from standard Dempster-Shafter combination theory represent the necessary steps to combine the sources.

$$m: 2^x \rightarrow [0,1]$$

Where 2X represents the power set of Θ or {0,{R},{Au},X} and where R and Au represent reliability and availability, respectively.

If m(Θ) is equal to 0, then the mass of the reliability and authenticity variables are represented as follows:

$$\sum_{A \in 2x} m(R) = 1$$

When combined, the joint mass of reliability and authenticity is represented by the following equation, where K is uncertainty:

$$m_{1,2}(R) = (m1 + m2)(R) = \frac{\sum_{R=Au \neq 0} m1(R)}{1 - K}$$

The value of m_1,2 (R) can be returned as a probabilistic value between 0 and 1. Thus m_1,2 (R) becomes the reliability-biased value for belief-based trust in identity. The same calculation can be performed with the authenticity as the lead variable with similar results.

Although useful for combining informational probabilities about a certain event, Dempster-Shafer theory is inadequate for the case of combining authenticity and reliability. Both authenticity and reliability are describing information about the same identity with its associated transactional context, which is a necessary precondition for the use of the Dempster-Shafer combination. However, it can be shown in the following sections that the two metrics analyze attributes differently. Thus they combination of the to would provide incomplete and possibly inaccurate information.

4.3 RELIABILITY

When a user offers an identity to a provider, a necessary component of trustworthiness is the ability to assess reliability. Just as trust exists between people, repetition of an appropriate behavior is necessary to continue building trust. In the example scenario, reliability equates to using the same credentials correctly and repeatedly. In the Belief Based Trust Model represented in algorithm 1, user-identity trustworthiness is computed from the perspective of the identity provider, and we focus on computing the reliability of the attributes presented.

Modeling reliability can be performed with probability distributions considering each interaction as a trial, and the overall user interaction with the identity repository considered as a sequence of trials [74]. In the example scenario, let x_i represent a match of each attribute presented by Bob and modeled as an independent random variable. The sequences of hits are treated as iid.

Taken together in a sequenced set, the attributes correspond to the transactional context from Bob using Alice's credentials and thus assuming her identity. The resource provider, ABC Industries, asks, "Can I trust that these presented attributes belong to Alice?" This question can be answered by calculating the probability that each presented attribute belongs to the identity that presented it. In this scenario, the presented attributes can be matched to an existing model of attributes for Alice's identity. Since such matching is possible, we can easily model reliability as a set of Bernoulli Trials. These random variables will either match an attribute in Alice's indexed identity or not.

Reliability as a metric is able to detect the repeated valid or invalid use of an identity. It is a useful measurement in a similar way to the use of reputation. Existing reputation based trust approaches assign a 1 or a 0 value to an actor based on a transaction [55]. This is insufficient when attempting to look deeper then just the actor.

For example an identity can use some attributes validly and others invalidly. Thus the reliability score can account for this mixed usage. A traditional reputation based methodology would consider the transaction invalid due to a single use of an attribute incorrectly. As Reiter said, some information is always valuable [9] regardless of the quality.

4.3.1 Reliability of Attributes

When considering reliability, it is crucial to look at attributes both as standalone items and as a grouping that forms an identity. A presented attribute can be considered reliable if its value is consistent with past values presented for that particular identity. In the example scenario, this means that the more that Alice asserts her FirstName attribute correctly, the reliability of that attribute type increases.

More formally, let Alice's FirstName attribute type stored with her identity be represented by A_1. Let the attribute presented as part of the transactional context be represented by A_0. Let each time Alice's identity attribute is presented be represented as a trial. Each trial should be considered independent. The independence assumption is required in order to correspond with how identities are used in the domain.

A user can conduct a transaction with the identity provider based on the type of protocol used by the identity provider. However, Alice's presented attribute could also be used simultaneously in another transaction with a malicious user. Thus, it is possible that many transactions in a row contain the same value for A_0. Furthermore, let us assume the values of A_0 to be a set of random variables. One possible trial outcome is the value of an attribute of type A_0 matching the value of type A_1. Thus, the state space S is defined as {0,1}, with 1 representing a successful match and 0 representing an unsuccessful match. We posit the following expression models the probability that presented Attribute A_0

matches the reference identity attribute across any number of trials *n* following binomial cumulative distribution function:

$$P_R = \sum_{i=0}^{|k|} {P_A}^i \binom{n}{i} (1 - P_A)^{n-i}$$

This equation provides a way to consider reliability across a certain *n* set of trials. The long run percentage of successful matches could indicate overall reliability, but this would not weigh the most recent transactions high enough. Using the binomial pdf above yields the ability to focus on most recent trials following a running estimate. This equation uses the probability of authenticity as the binominal probability, allowing P_R to increase greater across a lower number of trials when the attribute is more unique. This approach allows for modification of the independence assumption to provide for a more realistic view of the trials.

For example, Alice could be a long time customer of a certain identity provider. Several years ago, her identity could have been compromised when she incorrectly presented an attribute multiple times. If Equation 1x were to hold in this situation, then those probability estimates would be equally weighted with more recent negative matches. Thus, the independence assumption would be applied across a series of trials *n* in the Equation 1 asserting that each trial inside the most recent set of trials *n* is independent.

Let us call the probability in Equation 1 the recent reliability. The variable *n* will be fixed as a sliding window across the total lifespan of the attribute. Let *k-1* be the number of successful matches in the time window *n* where *k-1<n*. Accordingly, *k* will be the number of successful matches in the current trial, assuming that trial is successful.

Thus, the probability in Equation 1 will directly relate to the instantaneous reliability in the current transaction. This transaction timeline is represented in Figure 12.

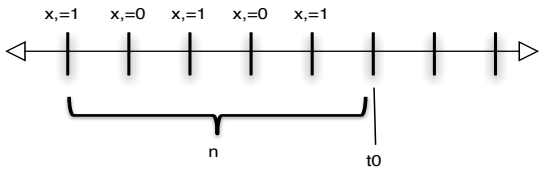

Figure 12: Example Transaction Timing

In Figure 12 the time of the current trial *t0* is used to calculate the recent reliability, where n=6 trial and k=4 successes (assuming that the trial is successful). Therefore, the recent reliability of rating at t0 is more relevant to the current context than if p had been calculated over the life of the attribute type. Following the approach from Equation 1 would yield similar results for reliability across the set of trials, illustrated by Figure 13.

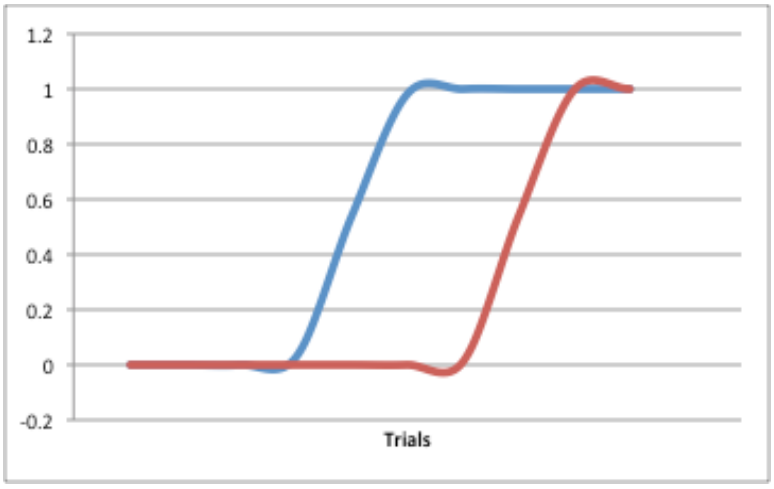

Figure 13: Trial Responses

As n is varied, the number of trials necessary to predict a positive match would increase as represented by the spacing between the blue and red curve. This behavior allows the calculation to be flexible enough to vary sensitivity as necessary for a particular set of domain conditions. The amount of time required to go from a negative match expectation to a positive match expectation is based on p, allowing the overall system reliability to provide bearing for the sensitivity of transition.

By delivering a measurement of the reliability of an attribute following equation 1, it is possible to get a view of attribute reliability across multiple transactions. Following the binomial approach still allows two outcomes reliable or not reliable but it allows the consideration of information some of which may be conflicting. Assessing conflicting identity information is necessary and sufficient for identification of a person or user[24, 26, 30].

4.3.2 Reliability of Identities

To define an identity as reliable, two conditions must be met: (1) the attribute types must be consistent across multiple transactions; and (2) attribute values must be the same across multiple transactions. Once reliability is asserted across any set of transactions, it can be used as a predictive metric. As previously discussed, a binomial distribution works well to represent reliability of an attribute. In terms of trust, the whole identity must be considered either trustworthy or not based on the grouping of attributes. In the example scenario, we expect Bob *not* to be a reliable user because he knowingly violated the provider's terms of service when he offered Alice's identity. The question, though, is to what degree Bob is considered unreliable. Following the attribute summary above, we can expect that reliability strictly depends on two factors: (1) the number of times an attribute matches represented as k; and (2) the size of the window of time in

measured in transactions represented as n. The below equation represents an approach for calculating identity reliability over time:

$$P_R = \sum_{i=0}^{|k|} {P_A}^i \binom{n}{i} (1 - P_A)^{n-i}$$

The above equation states that the reliability of the offered identity is a binomial function of n, k, and p. The variable n represents the total number of trials conducted or the number of times that the presented identity has been used to conduct transactions. The variable k represents the window size that is being considered for reliability. The window size represents the amount of transactions considered for whether or not a user is reliable. It can be sized high relatively to n if more precision is required but lower for less transactions and more recent data is required. The variable P_A represents the probability that the attributes used matched or the authenticity probability. P_A will be explained more in Section 4.4.

4.4 AUTHENTICITY

In addition to reliability, authenticity is a necessary to establish user-identity trustworthiness in online transactions. From this perspective, the study of identity resolution focuses on resolving identities and usage probabilistically based on a set of incomplete data. In effect, researchers are trying to match identity attributes with an index record held by the identity provider. In identity resolution, accuracy is a measure often used to demonstrate the success or failure of a particular algorithm [94]. Authenticity of attribute data can also mean s decision to trust was correct, if the index identity has matching attribute data [5]. Xiong[5] provides a confirmative view of trust

with attribute matching. This research proposes authenticity as a method to use attribute matching to establish a primary model for trust.

In the Trust Generation model (Algorithm 1) presented in this research, authenticity is based on the premise that a presented set of attributes corresponds with the person presenting them. Recall that in the example scenario, Alice gave Bob all of the information necessary to impersonate her. Therefore, we can assume that Alice would also share all of the necessary attributes to fully conduct an authentication transaction. We can also assume that all presented attributes correspond with the identity provider's indexed information about Alice's identity. In order to verify Alice's identity, the set of attributes Bob presents on behalf of Alice would need to be at least a subset of the attributes available in Alice's indexed identity.

$$\textit{Presented Attributes}\{A_0, A_1, \ldots, A_X\} \subseteq \textit{Index Identity}\{A_0, A_1, \ldots, A_Y\}$$

Let Au be the authenticity of presented attribute. Let A_x be the attribute being evaluated for authenticity. Authenticity will be characterized based on the amount of indexed identities that contain the presented attributes as a set. The below equation calculates the authenticity of attributes for an Identity I:

$$Au(I) = \frac{A_{correct}}{A_{total}}$$

Au is then represented as a real number from 0 to 1. The calculations for $A_{correct}$ and $A_{incorrect}$ are shown below.

$$A_{correct}(I) = \sum_{i=0}^{n} A_i * V(A_i)$$

$$A_{incorrect}(I) = 1 - A_{correct}(I)$$

An attribute is found to be matched or taking a value of 1 for A_i if the attribute value can be exactly matched to the golden record for the attribute of that particular identity. The value of $V(A_i)$ represents the number of attribute values matching the presented one within the golden set. Therefore the authenticity is both a function of how correctly the attribute is used and how many attributes are a direct match. Au is also inversely proportional to the number of attributes with the same value as illustrated by Figure 14.

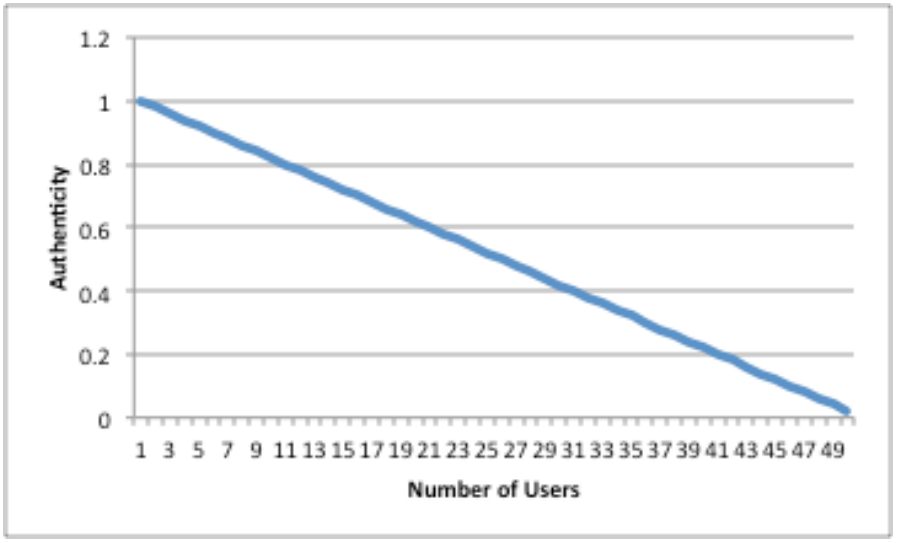

Figure 14: Authenticity vs. Number of Possible Occurrences of Value

In the example scenario, it is likely that Alice has a higher authenticity rating then Bob because her name is less common, and thus less likely to be used in the set of index identities. Au can only be values inside the real number set {0..1}. Thus, it can be combined with other attribute values using the previously described Dempster-Shafer approach to determine a belief metric. Identity providers can use the authenticity metric

to provide a comparative view of how unique a user's attributes are. It has the added ability to detect changes while at the same time detecting incorrectly used attributes. Existing trust mechanisms focus on the user as a whole including all identity attributes. They conduct a comparison to achieve a right or wrong answer and then provide that feedback to another trust engine. Authenticity has the ability to detect individual invalid attributes as well as those, which closely match other identities. If the identity were unknown or uncertain at the time of a transaction, it would be possible to use authenticity to produce a ranked order guess of the true identity based on the attribute values. More importantly, authenticity in this context has the ability to substantiate an assertion of identity through a probabilistic result.

4.5 EXPERIMENTS

The following experiments demonstrate that identity reliability and authenticity metrics defined by this research can enable more accurate trustworthiness assessments by an identity provider. If the attribute values vary significantly across multiple transactions, the reliability is too low and the provider loses trust in the identity. Additionally, a set of identities having a high number of instances with the same attribute values will lack uniqueness. Decreases in uniqueness (authenticity) decrease the ability of an identity provider to trust a given identity since others appear similar. The authenticity experiments combine attribute uniqueness numbers across all attribute types discussed in the previous section.

In order to demonstrate this trust-based approach a test set of identities obtained was analyzed by using an approach used by the U.S. Department of Homeland Security [92]. This dataset contained a sample of 50,000 identities that were representative of the population. The transaction data set is representative of the US population based on a few

properties. The first is the set of attribute types used: FirstName; LastName; Company; Address; City; County; State; ZIP; Phone; Email. The second is the cultural alignment of the names. Miller's team ensured that the names followed the cultural groups composition in society(circa 2009) 71% "American Names", 17% Anglo, Arabic, and Hispanic, ~9% Chinese Korean, Russian and Southwest Asian, ~3% French, German, Indian, Japanese and Vietnamese. The third property was to ensure that not all names were proper. Of the names chosen 1% were segmented into variants or nicknames.
The experiments were conducted to demonstrate reliability and authenticity responses when variables are varied (see Table 7).

Each experiment was configured to use the known set of identities from an identity provider. These were instantiated in a graph database in order to support efficient querying. These known identities will be referred to as golden record identities. Each time a transaction was conducted an identity with associated attributes was compared to the golden record in the database. Depending on the type of experiment, this approach was used to quickly and easily find attribute types with the same values, and it also keeps track of the number of times an attribute type and value are used across sets of identities.

Experiment	Attributes per Identity Transaction	Number of Transactions per Identity	Number of Values per Attribute Type	Number of Identities
Reliability Response to Random Attribute Types (4.5.1.1)	Varied [1-10]	10	Varied [5-500]	500
Reliability Response to Variation in Transaction Number (4.5.1.2)	Varied [1-10]	Varied [10-50]	50 [constant]	500
Authenticity Response to Variation in Transaction Number (4.5.2.1)	Varied [1-10]	Varied [10-50]	Varied [50-500]	500
Authenticity Response to Attribute Types (4.5.2.2)	Varied [1-10]	10	Varied [5-500]	500

Table 7: Reliability Experiments

Each experiment followed the computation of trust probability outlined in algorithm 1. First a transactional identity with a set of attributes is compared to existing ground truth identities. Once a set of possible ground truth identity nodes was returned, the remaining computations were performed on vertices connected by edges to that node. Each identity node was connected to a root index; this allowed one command to return the appropriate node (with up to 5 billion nodes) in less than 1s of database search time. This approach was able to efficiently find matches and return the necessary information. The maximum depth of traversal from the reference node was 3, with a breadth that depended on the number of stored attributes [95].

4.5.1 Reliability Experiments

The first metric, reliability of a user's identity, demonstrates the user is consistently who he or she claims to be, thereby increasing trustworthiness. In order to

prove that the reliability metric helps to indicate trustworthiness in an identity, we analyzed a set of transactions and varied the values provided for each attribute as well as the type of attributes used. For example, if Alice provides her address correctly in a string of multiple transactions, the reliability metric should be higher than if she provides it incorrectly in the string of transactions. The first experiment focused on determining how the reliability metric changed as each attribute's value varied between correct and incorrect for a specific identity over a period of transactions. Attribute types and attribute values per type were varied. The second reliability experiment built on the number of transactions in the first experiment by fixing the number of attributes used across all transactions and increasing the number of transactions.

4.5.1.1 Reliability response to random identity attributes.

The reliability metric was tested for responsiveness to changing identity attribute types and values. This experiment was conducted in a series of transactions that took a set of 10 identities and varied the attribute values and the attribute types. For example Alice would only communicate 3 personal attributes in one transaction while communicating 8 in another. Each transaction could have a random number of attributes that matched her ground truth identity. Reliability values for each of these identities were measured as each transaction was conducted.

It was hypothesized that if the number of attributes types per identity are varied randomly while the number of transactions per identity remains constant, then the reliability of each identity will be following a binomial increasing pattern so that reliability increases as the number of attribute types per identity increases following a pattern similar to the graph in figure 13. The experiment ran with a variation of 3 to 10

attributes per identity for each transaction. The number of same values per attribute type was varied from 5 to 500. Each identity conducted 10 transactions.

The hypothesis was verified. In Figure 15, each data point represents identity transactions with a certain number of attribute types. The reliability for each attribute was higher for those attribute types with larger numbers of similar values. Furthermore, when the amount of attributes per identity was low, the reliability was also low. In terms of user-identity trustworthiness, this is a useful indicator for calculating trust in an identity.

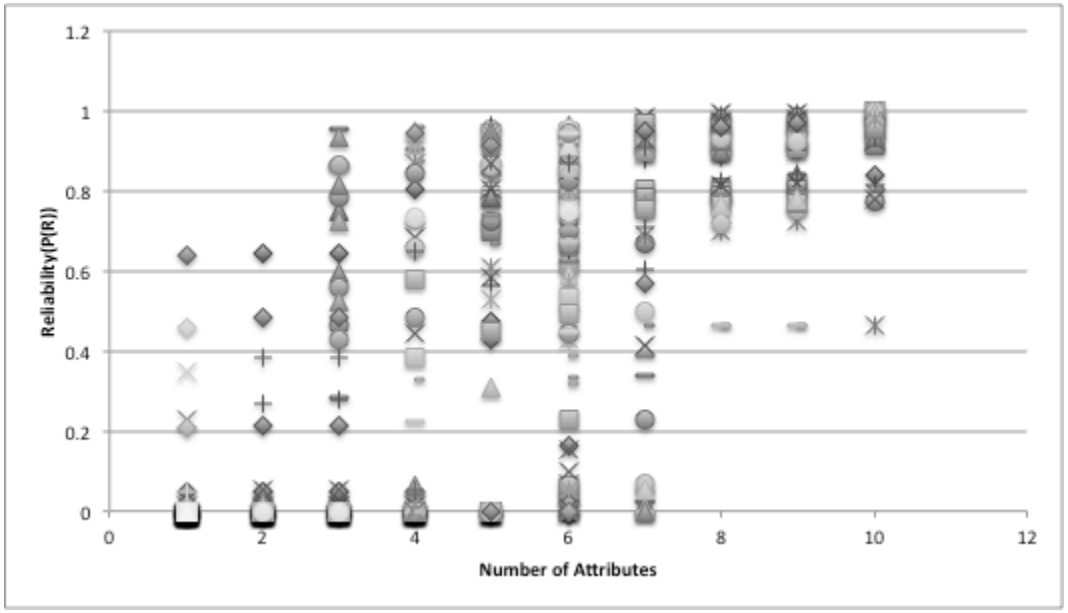

Figure 15: Reliability and Random Identity Attributes

These results can be applied to the example scenario by considering the case where Bob leverages Alice's credentials. If ABC Industries (the identity provider) had full knowledge of Alice's attributes and Bob only knew a few, such as her username and password, then Bob's reliability would be low if he were interrogated for more than just

the two attributes types that he knew. If Bob could only assert the two attributes correctly over a set of transactions, Alice's identity reliability score would not rise significantly. It is interesting to see how the reliability value for Alice's identity varies with more transactions perhaps at some point allowing her to conduct a transaction with the identity provider and reliably providing more of her attributes.

These results demonstrate that it is possible to gain a higher reliability by requesting more attributes. Practically, this is important to identity providers because they can choose to request a lot of attributes to achieve a high degree of trust in a transaction. Requesting more attributes can streamline the identity provider's business process. It already collects a significant amount of information about its users and being able to leverage it for a greater degree of trust could also save money. The Identity Provider would potentially be able to avoid investing in higher integrity attribute enrollment methods.

4.5.1.2 Reliability response to variation in transaction number.

Since an identity provider has to exist for longer than a fixed set of transactions, reliability must be tested across a large set of transactions. As users conduct more transactions, the reliability metric for each identity should increase due to the increase in time window assuming that the number of attribute types is fixed. This experiment also varied attribute values over a large set of transactions. Fixing the number of attribute types provided by a user should then ensure that reliability is calculated by numbers of matched attribute values. This situation would accurately reflect an identity provider like Google who might focus on a certain number of attributes types for a transaction but they have a wide population of users whose attribute values should be expected to vary significantly.

It was hypothesized that if the number of attributes types per identity remains constant while the number of transactions per identity is increased, then the reliability of each identity will increase as the number of transactions increases. The first experiment ran with the full complement of 10 attributes per identity for each transaction. Each identity conducted between 10 and 50 transactions.

Figure 16 shows a sample set of 10 identities from the 500 analyzed. Overall, 447 identities experienced an increase in the reliability over the set of 50 transactions. Identities initially had attributes values that didn't match ground truth, as the number of transactions increased each attribute still had a large number of correct values. This increase in the number of correct values directly led to an increase in the overall reliability of the specific identity being tested. In other words, an identity is more reliable when more of its attributes have been used more correctly across a larger number of transactions. The mean reliability for the set of Identities was 0.46. The mean transaction time to pass the mean reliability was 26 transactions. Surprisingly, the reliability numbers for this dataset passed .46 after 35 transaction runs. The large number of same attribute values for each transaction might explain the relatively low set of runs required for the reliability metric to approach 1. The jump in reliability is significant due to the choice of the window size. Since there are 50 transactions being considered, it takes approximately half of those to transition to a more reliable state. Thus it becomes important for a potential identity provider using this metric to choose the window size correctly for transactions measured.

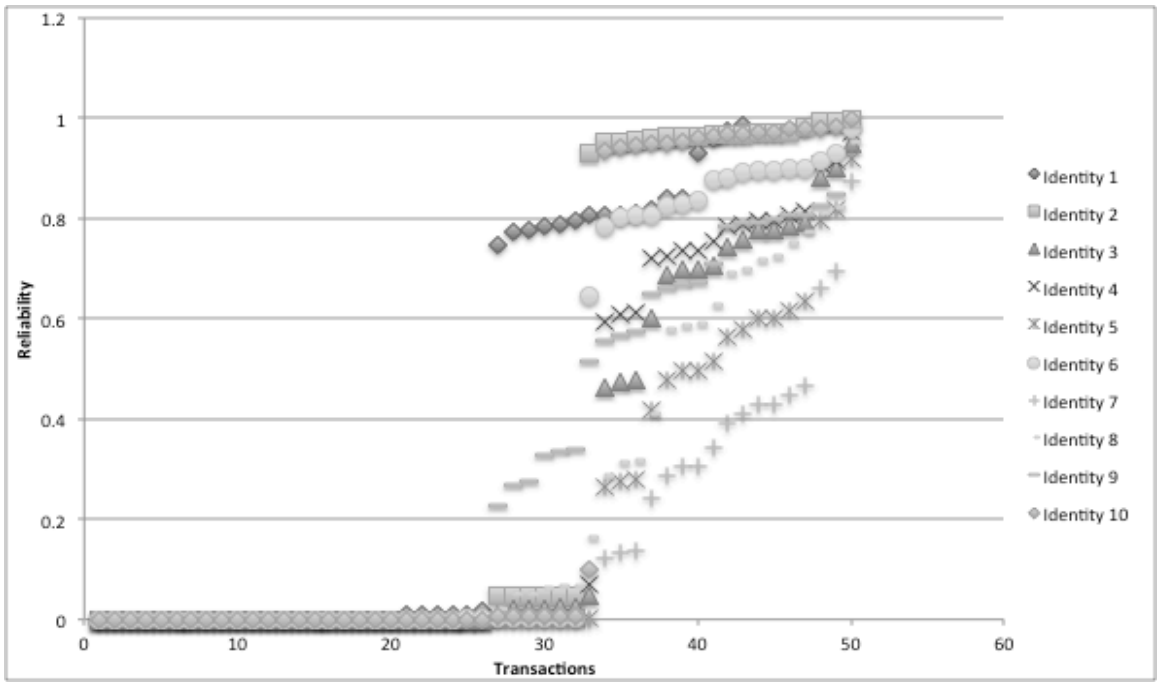

Figure 16: Reliability vs. Number of Transactions

Overall, 53 identities experienced a variable reliability instead of the expected increase. After revisiting the experimental configuration, we discovered those identities shared some characteristics including a larger number of duplicate values. For example, 31 identities shared the same first name. This finding suggests increasing the number of transactions may not impact reliability with a large number of similar attribute values across the identity set. These results are significant to example scenario if the reader considers Alice and Bob's transactions over a long period of time. If Bob is presented with situations where he entered nonmatching attribute values (attributes values that did not match those known as ground truth by the identity provider) for a long enough set of transactions, Alice's identity reliability value will decrease. This behavior could be useful to an identity provider as they could set a flag at a specified reliability value and

take some kind of remedial action in response to an unreliable identity, such as deactivating the account or forcing a password reset.

These results demonstrate that it is possible to gain a higher reliability by conducting more transactions. Practically, this is important to identity providers because they can choose to allow a lot of transactions to achieve a high degree of trust in certain scenarios where the user is consistently engaging in transactions. Often it can take several identity transactions for the user to get readable information from the service provider. From the identity provider's perspective this would be valuable to gain a higher degree of trust in the user being authenticated without requiring any additional infrastructure or algorithms.

4.5.2 Authenticity Experiments

An identity with more unique values is considered more authentic; these values can be leveraged to increase trustworthiness. For example, many people may have the name "Alice Jones," but fewer people named Alice Jones live on East Fifth Street, and even fewer have the same phone number. Thus, the more attributes types and more unique attribute values exist, the greater the authenticity metric in the user-identity trustworthiness calculation offered in this research. This metric would be easy to measure if all possible attribute values and types for a system of identities are known, and if the system is closed. But in reality, transactional systems are open, and identity providers must deal with how to measure authenticity with incomplete information about the identities that will conduct transactions.

4.5.2.1 Authenticity response to variation in transactions by attribute.

Remembering the definition from section 4.4, authenticity will be characterized based on the amount of indexed identities that contain the presented attributes as a set.

Since an identity provider has to service more than a fixed set of transactions, the authenticity must be tested across a large set of transactions. As users conduct more transactions, the authenticity metric for each identity should increase due to increased attribute types matched to a single identity.

For example, as transactions increase, Alice Jones will supply more information about herself and the authenticity of her identity will increase as more attribute types are used. This experiment fixed the number of attribute types with a variation in user attribute values over a large set of transactions. Fixing the number of attribute types provided by a user should then ensure that the authenticity would reach an upper limit. This situation would accurately reflect an identity provider like Google, who might focus on a certain number of attributes types for a transaction, but also have a wide population of users whose attribute values should be expected to vary significantly. Therefore, if an identity is used across several transactions, and attributes types are fixed each time while the attribute value is held varied, the identity's authenticity should increase as the experiment proceeds. This experiment was conducted with the further assumption that a user conducting it wasn't malicious and had full knowledge of an identity.

It was hypothesized that if the amount of attributes types per identity remains constant, the authenticity of each identity will increase linearly based on the number of transactions. The experiment was run 5 times with a different number of attributes each time. The number of values for each attribute type was varied across the set of transactions. All identities conducted transactions until they reached the highest authenticity score possible. The number of attribute values per attribute type varied across the set of transactions.

This hypothesis was verified providing proof that authenticity is dependent on both number of attribute types and transactions. The authenticity for each attribute type

depended only the number of transactions and not the number of duplicate attribute values (see Figure 17). Each identity used in a transaction was measured. Figure 10 displays 5 sample results from the experiment. It took about 50 runs for the Identity to reach maximum value with 5 attribute, and 192 with 1 attribute. The findings support the hypothesis that authenticity increases as the number of transactions increase. One notable behavior is that it takes a significant number of transactions to reach peak authenticity with only one attribute.

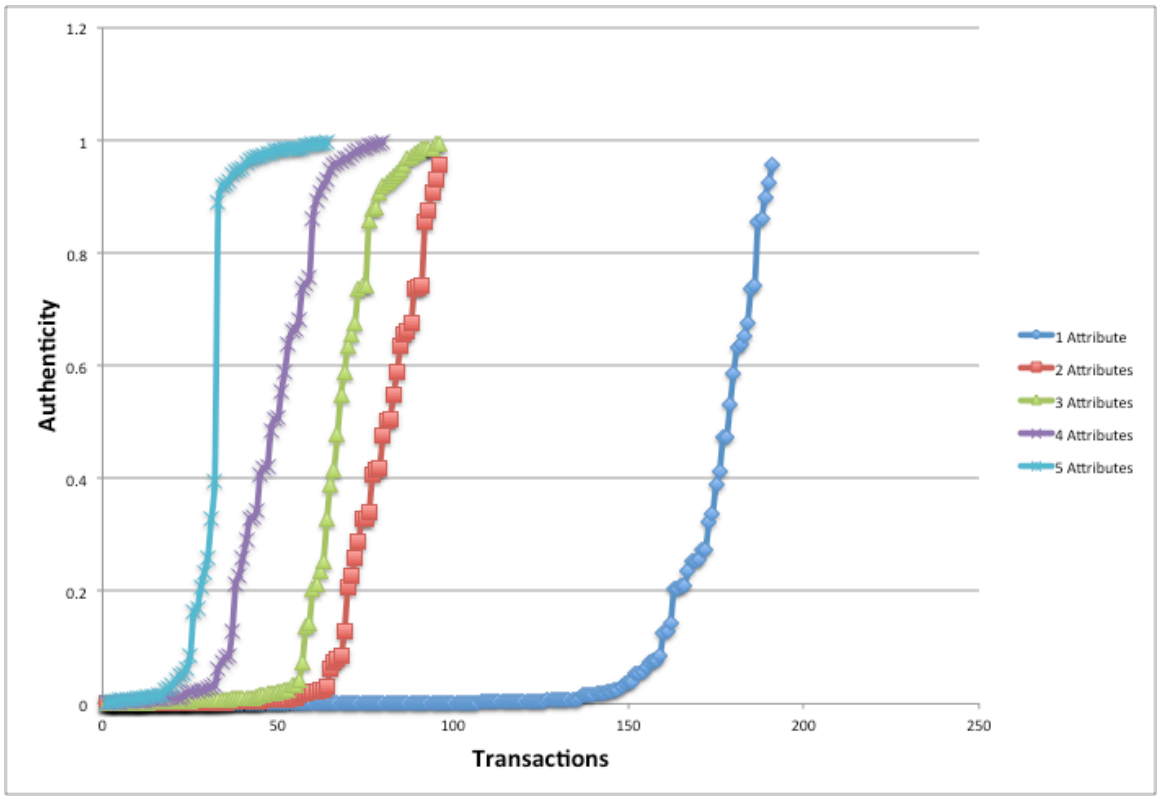

Figure 17: Authenticity and Number of Transactions

This type of experiment is useful in the context of scenario 3 because if Alice had been conducting transactions with ABC's identity provider, she would constantly be providing more identifying attributes about herself thus increasing her identity's

authenticity. The expected response of this metric would be to decrease the authenticity as the amount of correct identity attributes were decreased. The decrease would support a case where an account was created for a fake identity and the user conducting the transactions didn't have the full set of attributes needed by the identity provider.

An identity provider can find the authenticity metric useful due to the relatively short amount of time necessary to converge on a high authenticity value. Once the identity has achieved authenticity, due to the number of transactions conducted, it stays high. It can be asserted that as the number of valid transactions decreases the value falls just as fast. This property is more useful because an attacker might take a few times to correctly guess attributes of the user's identity and thus produce a low reliability value. The next section will discuss what happens when the number of valid attributes is varied.

4.5.2.2 Authenticity response to variations in the number of attributes.

Authenticity will be characterized based on the amount of indexed identities that contain the presented attributes as a set. Building on the results from the previous experiment, it is useful to understand how authenticity varies with respect to number of attributes. In order to demonstrate this a certain number of identities can be studied with respect to their attribute usage. This should clearly indicate whether number of attributes affect authenticity. One would expect that as Alice Jones will supply more information about herself and the authenticity of her identity will increase as more attribute types are used correctly. This experiment chose a fixed number of identities varying the number of attribute types used to conduct transactions. Finally, this experiment was conducted with the further assumption that a user conducting it wasn't malicious and had full knowledge of an identity.

It was hypothesized that the authenticity of an identity will increase as the number of attribute types used for transactions increases. The first experiment with 10 identities and varied the number of attribute types used. All identities conducted 50 transactions. The number of attribute values per attribute type varied across the set of transactions.

This hypothesis was verified substantiating the view that the more attributes used correctly, the higher the authenticity value. The authenticity for each identity depended only on the number of attribute types and not the number of duplicate attribute values (see Figure 18). Each identity used in a transaction was measured across the full set of transactions. Figure 18 displays 10 identities from the experiment.

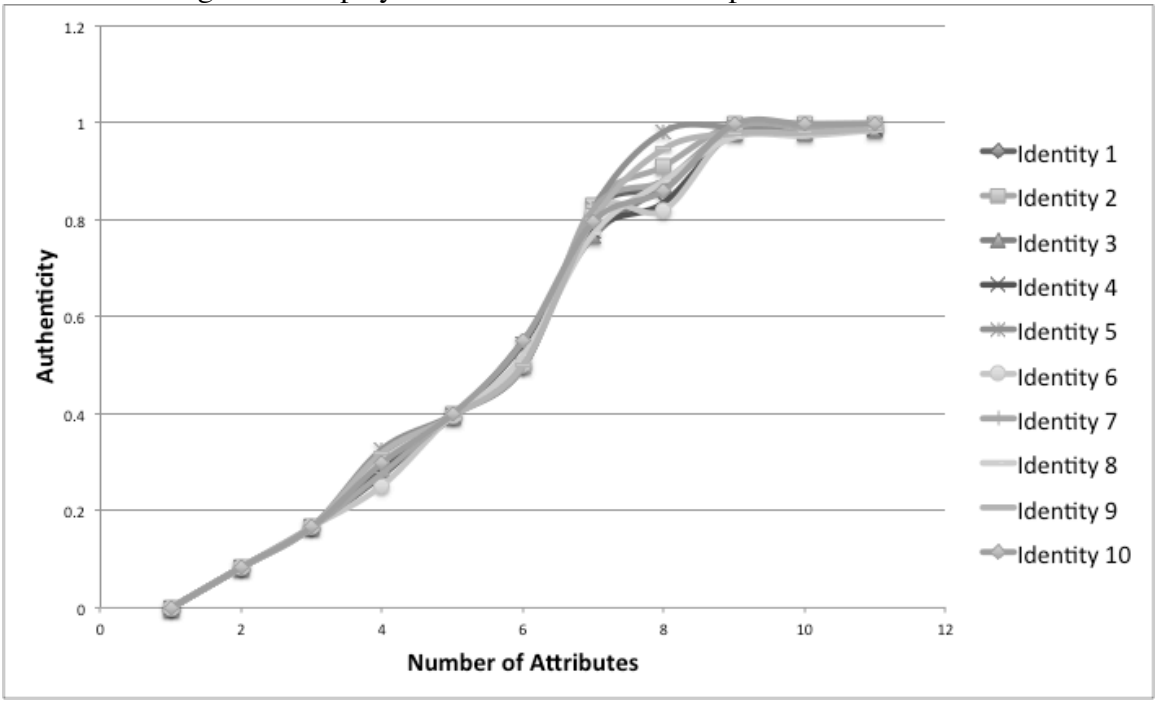

Figure 18: Authenticity vs. Number of Attributes.

As the experiment progressed, each identity eventually matched enough attributes from the transactions to increase the identity authenticity. Authenticity values reached their peak at about 8 correct attributes. This peak is representative of the transaction set,

which had the same amount of valid and invalid transactions for the first 7 attributes. As the set of attributes went higher then 8, the validity of the attributes stayed high.

An identity provider can find the authenticity metric due to the relatively linear increase in authenticity vs. the number of attributes. This property is more useful because an identity provider decide to increase attributes requested and have an expected return authenticity. A simple comparison of values can allow the identity provider a view into how may return attributes are invalid.

4.5.3 Reliability with Attribute Preference

The first metric, reliability of a user's identity, demonstrates the user is consistently who he or she claims to be. For example, if a person uses identity attributes that are more likely to be compromised to identify him or herself, then the reliability of his or her identity should decrease. In order to prove that the attribute used is important to the calculation of the reliability metric, a set of transactions was analyzed that varied the values provided for each attribute, as well as the type of attributes used. However, if the attribute has a higher P_C, then the reliability number should decrease proportionally. Thus, equation 3 can be modified to show improved reliability as shown below:

$$\textit{Equation 4}\quad ImpRel(I_0) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} P_C(A_k) * P_R(A_k)}{nm}$$

Improved reliability adds the Probability of Compromise as a multiplier to the existing probability of matching the particular attribute. Returning to the example, if Alice provides her address correctly in a string of transactions, the reliability metric would be higher than if she provides it incorrectly in the string of transactions. However,

if she provides her address incorrectly in a string of transactions, and also provides her SSN correctly in a string of transactions, the reliability metric should increase more than proportionally. Leveraging the results from the first reliability experiment set, a comparison was made between calculating reliability with a random set of attributes and specific types.

It was hypothesized that if the number of attributes types per identity is varied randomly while the number of transactions per identity remains constant, then the reliability of each identity will be higher as the Probability of Compromise for an attribute decreases. This experiment ran with from 1 to 10 attributes per identity for each transaction. The number of values for each attribute type was varied across the set of transactions. All identities conducted 50 transactions. The number of attribute values per attribute type varied across the set of transactions. This experiment was run three times. The first time randomly selected the attribute types, the second time selected the attributes with the top 10 highest Probability for Compromise, and the third time selected the attributes with the bottom 10 highest Probability for Compromise.

This hypothesis was verified demonstrating that Improved Reliability was indeed affected by the number of attributes by type. The attributes with the higher probability compromise scores yielded lower reliability scores. The attributes with the lower probability scores yielded higher reliability scores. Figure 19 shows the results.

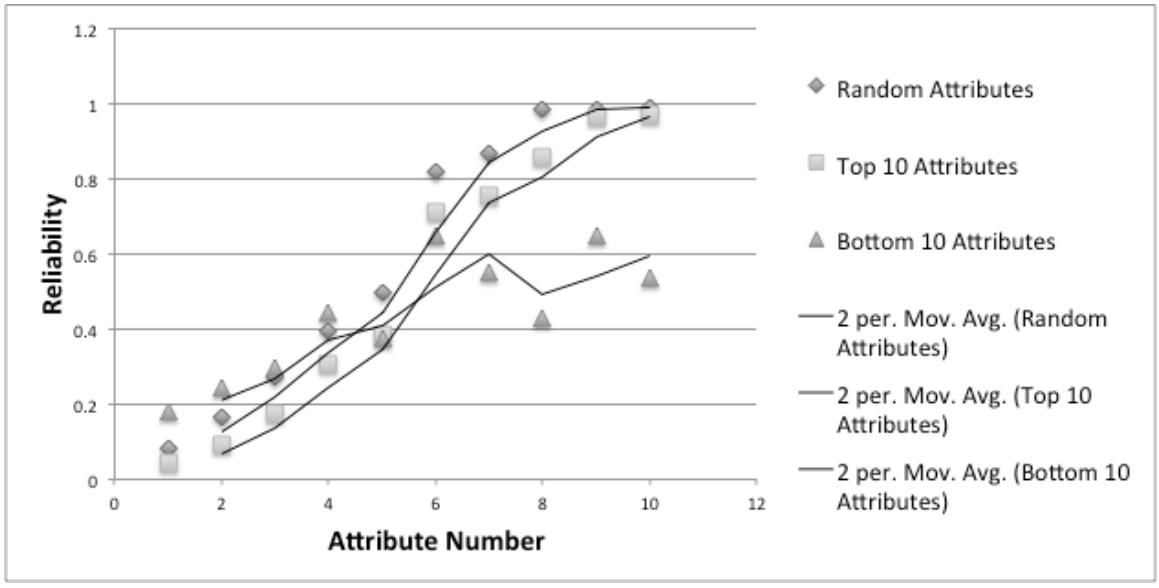

Figure 19: Improved Reliability Results

The Improved Reliability Metric yielded statistically significant results with a mean difference of .021999 in a t-test with a standard deviation of .01021 and 9 degrees of freedom. The conclusion is that the improved reliability metric is more useful then the standalone reliability metric due to the addition of P_C in the calculation.

However the Improved Reliability metric shows insufficient for prediction of reliability based on types of attributes used. Ideally, the metric would return a statistically significant higher value for the attributes, which are less likely to be compromised. Unfortunately there is not a significant enough change to show that the attributes used have bearing on the improved reliability metric. This leads the dissertation to try a different approach.

Chapter 5: Identity Risk

Trust in an identity is not enough to answer the question of whether a resource provider should grant access or authorization to a user. Basic trust in identity provides a probability that the identity is correct. However, there still is not a method for interpreting this probability as a useful assessment for access control decisions. Improved Reliability proved insufficient to use attribute compromise probabilities in the assessment of validity of an identity. Modifying Kim's approach [99] to consider risk as a decision to trust based on 3 factors seems an appropriate method to try.

Recall that RQ3 asked: *Is the risk posed by an identity a reliable predictor of the validity of the identity?* This question involves reliability, improved reliability and authenticity measures for attributes, and identities. According to Kim, et al [99], trust and its antecedents including reliability can be an accurate predictor of risk in an e-commerce environment. Therefore it makes sense to leverage these measures to determine the risk posed by the identity in the following scenario.

5.1 Example Scenario for Risk Calculation

Based on Kim's principle that trust affects perceived risk in an identity [99], the risk in a user gaining access to resources should be based on all available information about that user. To illustrate this concept, let us revisit the hypothetical example scenario. Once again, Alice is collaborating with Bob and Mavis intercepts their communications. This time, however, ABC Industries sets up defenses against the faulty use of credentials, challenging the authenticator with a knowledge-based tool that requires another attribute of the end user to verify authenticity. Initially, this seems like a reasonable solution to the reliability and authenticity problems presented in the previous

chapter. Unfortunately, the attribute has a Probability of Compromise that is not 0, and to compensate, ABC has to understand the risks an identity can pose to their infrastructure.

Such a scenario seems to be commonplace in online transactions, and multiple solutions have been developed to increase user-identity trustworthiness based on credentials including encryption, multiple factor tokens, and secured networks. However, these solutions fail to solve the fundamental problem of authenticating a user's identity: How can we ensure that a user is who he or she says that they are? Therefore, this research also makes the following assumptions in creating a new method for calculating risk from a provider's perspective:

- Risk is dependent not only on the trustworthiness of the credential used but also on the attributes used to create that credential.
- Risk is a useful tool to aid in authentication decisions.

Based on these assumptions, this research posits that users can be identified with greater confidence by calculating risk using trustworthiness techniques and existing information about attributes.

5.3 IDENTITY RISK CALCULATION

The following section proposes an approach for calculating risk for a user-identity. The attributes presented with the identity are combined with the trust value gained from the reliability and authenticity metrics to form an identity risk assessment. The risk model is based on the identity models discussed in previous chapters. The definition of identity remains consistent and risk is calculated for a specific identity. In general, the risk model follows the pattern presented in Figure 20.

The risk calculation determines the probability that users presenting identities are who they say they are, based on all available information. This is a function of their

reliability, authenticity, and the Probability of Compromise of the attributes that they present. A key design feature of the two trust metrics is that they are expressed as probabilities. Viewing the problem as a set of probabilities allows the ability to exceed the Dempster Shafer and Improved Reliability approaches. In order to combine these metrics, a few approaches for combination of probability were investigated. The approach must reasonably combine potential conditionally dependent data.

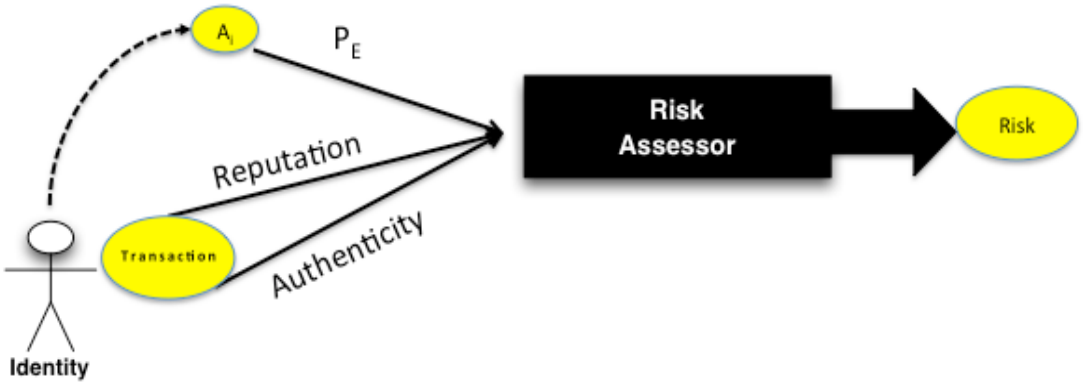

Figure 20: Risk Assessment Approach

One approach is to use a Bayesian Network [38, 39, 75] to determine the combine modalities of these component probabilities. For example, Bissiri's approach uses a Bernoulli combination technique to combine disparate identity information [38]. Muncaster shows that a Bayesian approach can encode a multivariate set of probabilities for identity attributes such as biometric data [75]. The authors assume an unknown posterior distribution to enable them to fine tune their inferences. These techniques are both useful in different ways and can be leveraged for the combinations of trust metrics with the P_C for each attribute.

The following model details the combined probability approach to computing risk. Let the compromise Probability, P_C for an attribute be independent from the

reliability and authenticity metrics for identity. The conditional independence assumption is valid because P_C is the Probability of Compromise for a specific attribute. P_C is derived directly from the attribute graph that details connections between the sets of attributes. However, P_R, and P_A are possibly dependent. Therefore, they will be treated as conditionally dependent for the sake of the computation.

The behavior can be modeled by using a stochastic process. Since reliability is a way of judging the trust placed in an identity based on its use of attributes, P_R for I_0 would be combined with P_C (I_0) to show the risk for that identity. Figure 21 represents the stochastic process that defines risk for an identity.

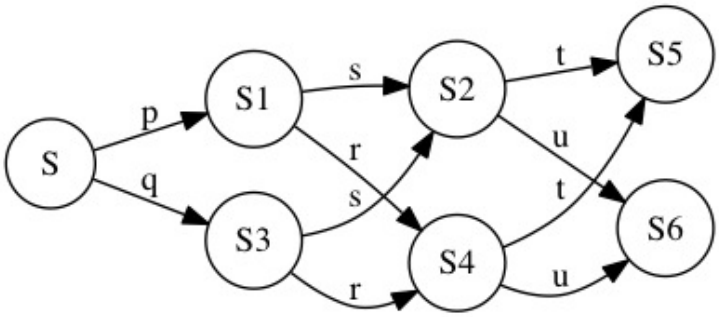

Figure 21: Risk Stochastic Process

Let S be the initial state of an Identity during a transaction. Let S1 be a choice to rely on the identity and S3 be a choice to determine its unreliability. S can transition to S1 with a probability *p*, which is 1-P_R. S can also transition to S3 with a probability of q, which is P_R. A transition to one of the states describes a higher probability to be reliable. This stochastic model, however, makes a decision about the reliability of the user in the context of Probability of Compromise of the attributes used for that particular transaction. S1 can transition to S2 with a probability of r or S4 with a probability of s. S3 can

transition to S4 with a probability of r or S2 with a probability of s. The variable r denotes a compromise Probability P_C and s denotes 1-P_C. The r transition is made then the attributes in the identity $P_C(I)$ are considered not compromised. S5 and S6 follow the same pattern except to consider authenticity with respect to reliability. If a t transition is made then the identity is shown to be authentic whereas with a u transition the identity is not authentic.

Since P_C generally reflects geometric distribution, the threshold for determination is made based on the mean. After parametric regression smoothed distribution was obtained (see Figure 22).

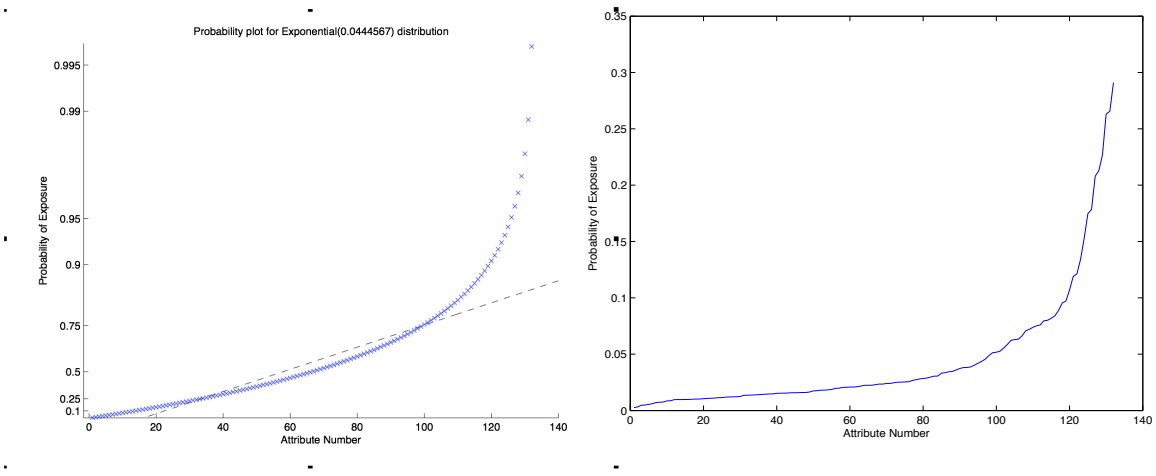

Figure 22: Fitted Probability of Compromise Distribution

The PDF of the fitted function is $y = e^{-.0445x}$ with a mean=.045. Thus, the decision to move to S2 or S4 is based on whether the attributes are considered compromised based on the mean value. This process is provably stochastic as it is composed of two transitions that sum to 1. The system is a closed communicating class and positively recurrent.

This stochastic model is valid for a single transaction. When repeated across multiple transactions, this stochastic process can be generalized as a Bayesian Network for computational purposes. Following the process in Muncaster [75], this stochastic process is decomposed to a simple Bayesian Network shown in Figure 21.

The network can be optimized across a large set of transactions to provide a risk calculation for each transaction. The Bayesian network is also extended to include the authenticity probability. Extending the network provides the ability to consider the amount of same valued attributes within an attribute type. An identity provider such as USFA in the example scenario would be able to make an informed judgment about the risk posed to their network, as each transaction occurs based on all of the available metrics. The joint probability function that defines the overall risk probability is:

$$P = (Pe, Pa, \Pr) = P(Pe|Pa, \Pr) * P(Pa|Pe) * Pe$$

The most useful questions for calculating a valid risk, along with their associated equations, are:

1. What is the probability that an attribute asserting the identity is compromised if the identity is not reliable?

$$P(Pe = T|Pr = F) = \frac{P(Pe, \Pr)}{P(\Pr)}$$

2. What is the probability that an attribute asserting the identity is compromised if the identity is not authentic?

$$P(Pe = T|Pa = F) = \frac{P(Pe, \mathrm{Pr})}{P(\mathrm{Pa})}$$

3. What is the probability that an attribute asserting the identity is compromised if the identity is not authentic and not reliable?

$$P(Pe = T|Pa = F, Pr = F) = \frac{P(Pa, Pe, \mathrm{Pr}) * P(Pe|\,\mathrm{Pr})}{P(\mathrm{Pe}\,|Pa)}$$

4. What is the probability that an attribute asserting the identity is compromised if the identity is authentic and not reliable?

$$P(Pe = T|Pa = T, Pr = F) = \frac{P(Pa, Pe, \mathrm{Pr}) * P(Pe|\,\mathrm{Pr})}{P(\mathrm{Pe}\,|Pa)}$$

5. What is the probability that an attribute asserting the identity is compromised if the identity is not authentic and reliable?

$$P(Pe = T|Pa = F, Pr = T) = \frac{P(Pa, Pe, \mathrm{Pr}) * P(Pe|\,\mathrm{Pr})}{P(\mathrm{Pe}\,|Pa)}$$

Each of these equations can be summed to produce the likelihood that an identity is not accurate. The likelihood that an identity is not accurate reflects a value that shows the risk for the transaction. This risk value is useful for detecting invalid transactions. An invalid transaction is defined as one that a user submits as an incorrect value for an attribute type.

There could be several reasons for an invalid transaction such as a person forgetting their password, or a user maliciously attempting to use another identity. Since

the risk model has to predict whether an identity is valid or invalid, the probabilities calculated must be divided up into a binomial distribution to choose validity. The Probity model chooses a mean that is represented by P_R^* across both types of transactions if the transaction has a valid probability that falls above the mean, then that is a predictor that the user will be valid. The probit model is represented by the below equation where x is the value of an individual data point for risk:

$$\text{Probability}(P_R=1|x)=\text{Probability}(P_R^*>0|x)$$

Since the valid and invalid probability sum to 1, the invalid probability will automatically be below the P_R^*. If the transaction has a valid probability that falls below the P_R^* and the invalid probability is above the mean that is a predictor that the identity will be invalid. A graphical view of valid and invalid probabilities graphed against attribute type. Figure 23 represents a calibrated view of data that was used to create the improved reliability metric. It is obvious that at above 6 attributes, the risk model from figure 21 is almost perfect at prediction with prediction thresholds set very small around 0 and 1. With 6 attributes or less, the prediction threshold sizes need to increase a little bit to make precise predictions. There is some error in the model but the overall accuracy of prediction is 72.6%. This is significant as it can be used as an improvement to the way that service providers conduct transactions. For example, if service providers could predict with a high degree of accuracy which of the identities they let through were invalid, they could supplement their authorization mechanisms. These mechanisms could adapt dynamically to changing conditions rather then just require a set of attributes to gain access.

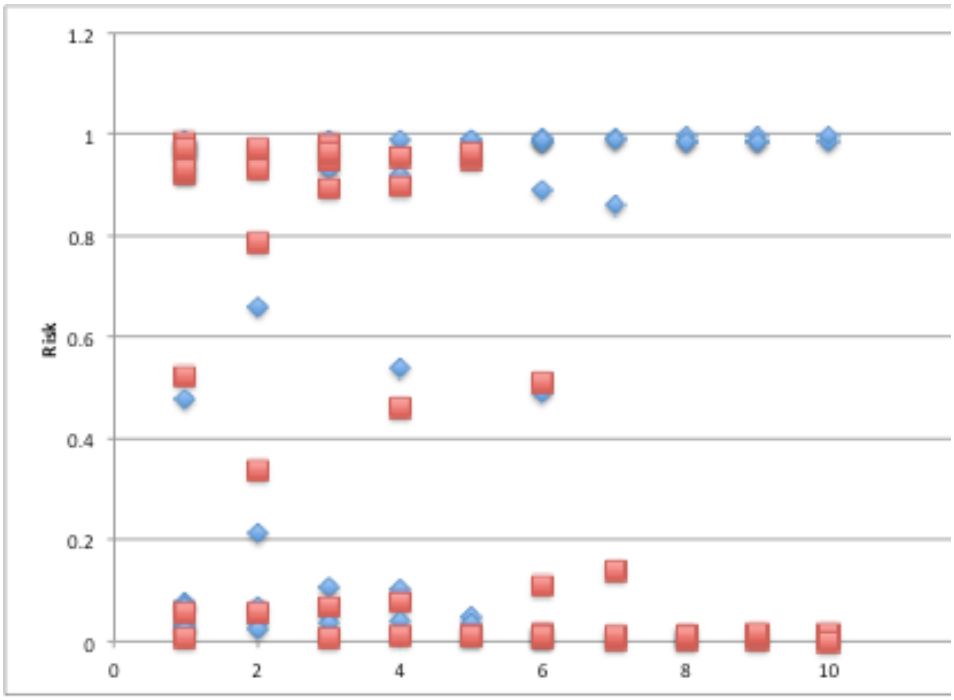

Figure 23: Risk Probability vs. Attribute Type

The domain of information retrieval works best to evaluate this type of approach. Table 8 shows the measurement approach for the results of the risk probability calculation. The first applicable metric applicable is the precision. A true positive is represented by an identity shown as valid based on its probability falling above the mean in the Probit model..

		Gold Record	
		True	FALSE
Outcome of Test	Valid	True Positive	False Negative
	Invalid	False Positive	True Negative

Table 8: Test Outcomes

A false positive occurs when a record is assumed valid based on the validity being above the mean in the Probit model but the identity record doesn't match.

$$Precision = \frac{Tp}{Tp + Fp}$$

Building on the precision measure, mean average precision (MAP) can be used across the dataset to determine the relative utility of the risk measurement approach. MAP is derived by first calculating an average precision value for each query in the dataset, and then averaging across the number of queries.

$$MAP = \frac{\sum_{i=1}^{n} P(i)}{n}$$

Another measure to determine the utility of the risk calculation is accuracy. In order to understand accuracy, another term must be defined. A true negative occurs when the record is chosen to be invalid based on being below the Probit model mean.

$$Accuracy = \frac{Tp + Tn}{\sum_{i=1}^{n} Transation(i)}$$

5.5 EXPERIMENTS

The following experiments allow for better reasoning about how to combine belief and reputation modalities to best develop an assessment risk metric associated with a given identity. The specific measure will be described in terms of precision, accuracy, and mean precision. The overall risk calculation will be realistically computed based on the example scenario. The expectation is with all metrics, the risk model should be a

good predictor of an invalid transaction. Table 9 below breaks down the experiments with associated variables. Each experiment ran with 800 transactions.

Experiment	Section	Measurements	Number of attribute types
1	5.5.1.1	Risk Accuracy combining Attribute Compromise and Authenticity	Varied [1-10]
	5.5.1.2	Risk Precision combining Reliability and Attribute Compromise	Varied [1-10]
2	5.5.2.1	Risk Accuracy combining Attribute Compromise and Reliability	Varied [1-10]
	5.5.2.2	Risk Precision combining Reliability and Attribute Compromise	Varied [1-10]
3	5.5.3.1	Risk Accuracy combining 5 valid risk probabilities	Varied [1-10]
	5.5.3.2	Risk Precision combining 5 valid risk probabilities	Varied [1-10]

Table 9: Experimental Approach

5.5.1 Usage of Authenticity to determine Risk

The risk metric was tested for variations in accuracy for changing numbers of attribute types. This experiment ran a set of transactions in order to calculate risk using only Probability of Compromise of all attributes used by an identity and authenticity. This experiment was conducted in a series of transactions that took a set of 16 identities and varied the number of attribute types. For example, Alice Jones communicates 3 personal attributes in one transaction and 8 in another. Each transaction could have a random number of attributes matching her ground truth identity. Risk values were

measured for each of these identities as each transaction was conducted following the same procedures in Chapter 4.

5.5.1.1 Risk Accuracy Response to Variations in Number of Attribute Types

It was hypothesized that if the number of attributes types per identity are varied randomly while the number of transactions per identity remains constant, then the risk accuracy of each identity increases as the number of attribute types per identity increases. The experiment ran with a variation of 1 to 10 attributes per identity for each transaction. The number of same values per attribute type was set at 20. Each of 16 identities conducted 50 transactions for a total set of 800 transactions.

The hypothesis was verified demonstrating that there was variation of risk based on the number of attribute types used. The data collected is represented in Figure 24 below. In Figure 24, each data point represents mean accuracy for an attribute type. Accuracy clearly increases as the number of attribute types used increases. When the amount of attributes per identity is low, the authenticity is also low. This is a possible factor of the sharp rise in accuracy.

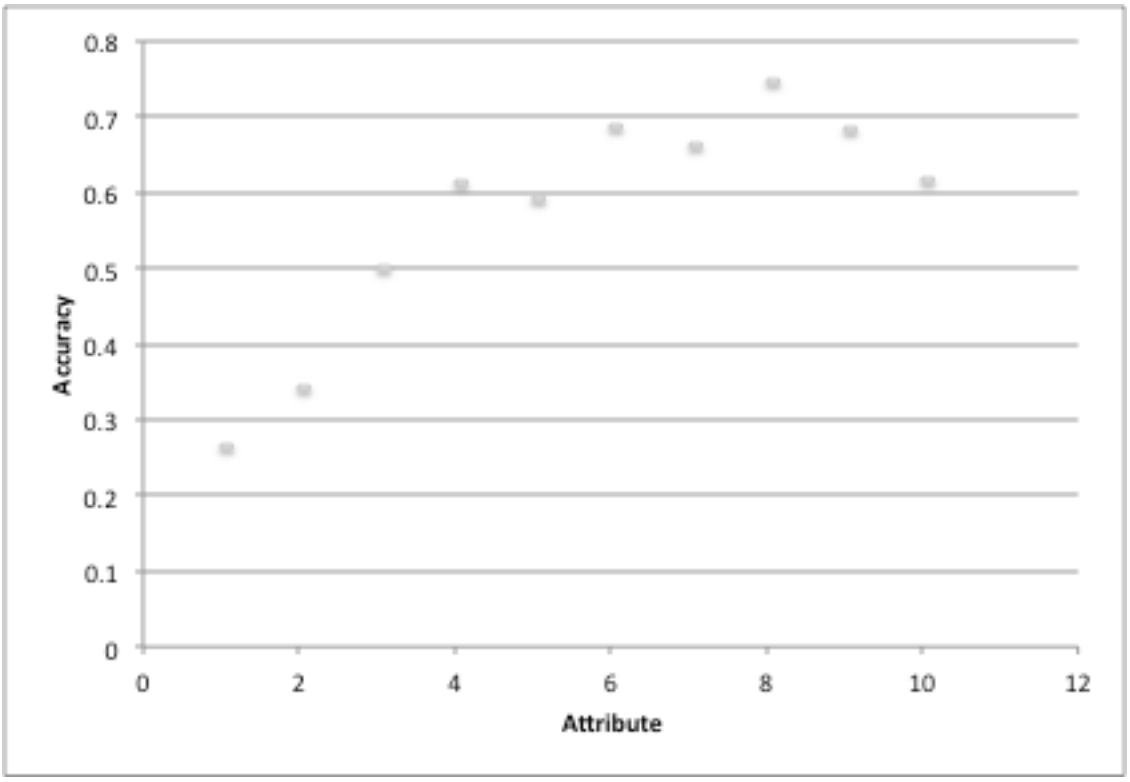

Figure 24: Accuracy Response by Attribute Type

The low authenticity values for a small amount of attribute types is expected because an identity proves itself more authentic as it uses larger amounts of attribute types to identify itself. Put differently, the fewer identity attributes are used, the higher the risk of that user if risk depends on authenticity. Clearly, the attribute probabilities of compromise are used because they tend to level the accuracy of the risk metric. For example, it can be asserted the relationship is not fully linear because attributes 9 and 10 have a higher Probability of Compromise. It is expected that since the accuracy of risk prediction generally increases as the number of attributes increase, that a service provider would request more attributes be used in order to get a higher risk accuracy.

5.5.1.2 Risk Precision Response to Variations in Number of Attribute Values per Type

This section focused on combining Probability of Compromise of all attributes used by an identity and authenticity. The experiment was conducted in a series of transactions that took a set of between 16 identities and varied the number of attribute types. Risk precision values were measured for each of these identities as each transaction was conducted.

It was hypothesized that if the number of attributes values per identity is varied randomly while the number of transactions per identity remains constant, then the risk precision of each identity increases as the number of attribute types per identity increases. The experiment ran with a variation of 1 to 10 attributes per identity for each transaction. The number of same values per attribute type was set at 20. Each identity conducted 50 transactions.

Again, the hypothesis was verified. The data collected is represented in Figure 25. In Figure 25, each diamond represents the precision for identity transactions with a certain number of attribute types. The precision is higher for those attribute types with larger numbers of similar values. Furthermore, when the amount of attributes per identity is low, the precision is also low.

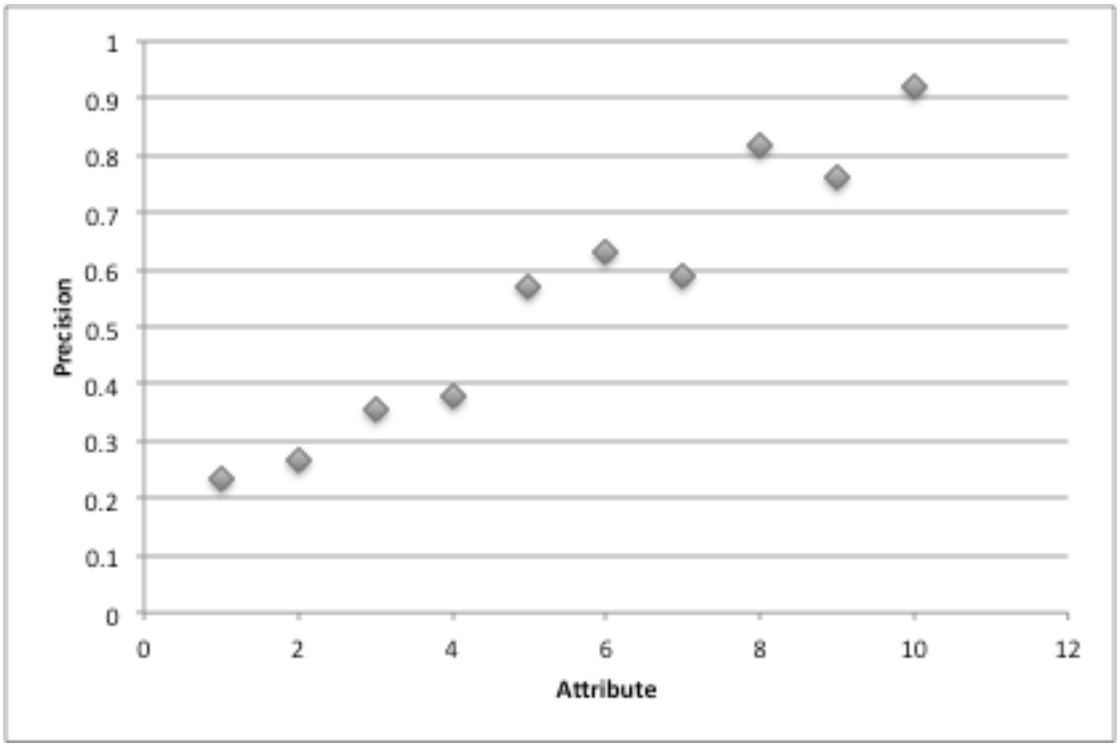

Figure 25: Precision Response by Attribute Type

The low precision values for a small amount of attribute types are expected, because an identity proves itself more authentic as it uses larger amounts of attribute types. These low precision values are primarily due to the preference of the authenticity values towards detection of true negatives than detecting true positives. Service providers should be concerned about a low precision at low levels of attribute usage. Even with a somewhat high accuracy number an attribute provider who uses up to 4 attributes still has to deal with the potentially significant error introduced due to the low precision valid under. 5.

5.5.2 Usage of Reliability to Determine Risk

This section was intended to elucidate the changes in risk based on only Reliability and Probability of Compromise of all attributes used by an identity. The

experiment was conducted in a series of 800 transactions that used a set of 16 identities and varied the number of attribute types. The variation in attribute types was performed in order to understand the levels of risk posed by using different amounts and types of attributes. Risk values were measured for each of these identities as each transaction was conducted following the same procedures in Chapter 4.

5.5.2.1 Risk Accuracy Response to Variations in Number of Attribute Values per Type

It was hypothesized that if the number of attribute values per identity is varied randomly while the number of transactions per identity remains constant, then the risk accuracy of each identity will increase as the number of attribute types per identity increases. The experiment ran with a variation of 1 to 10 attributes per identity for each transaction. The number of same values per attribute type was set at 20. Each identity conducted 50 transactions.

This hypothesis was also verified showed that there was variation of attributes used in the calculation based on the reliability metric. The data collected is represented in Figure 26. In Figure 26, each data point represents mean accuracy for an attribute type. Accuracy clearly increases as the number of attribute types used increases. When the amount of attributes per identity is low, the authenticity is also low.

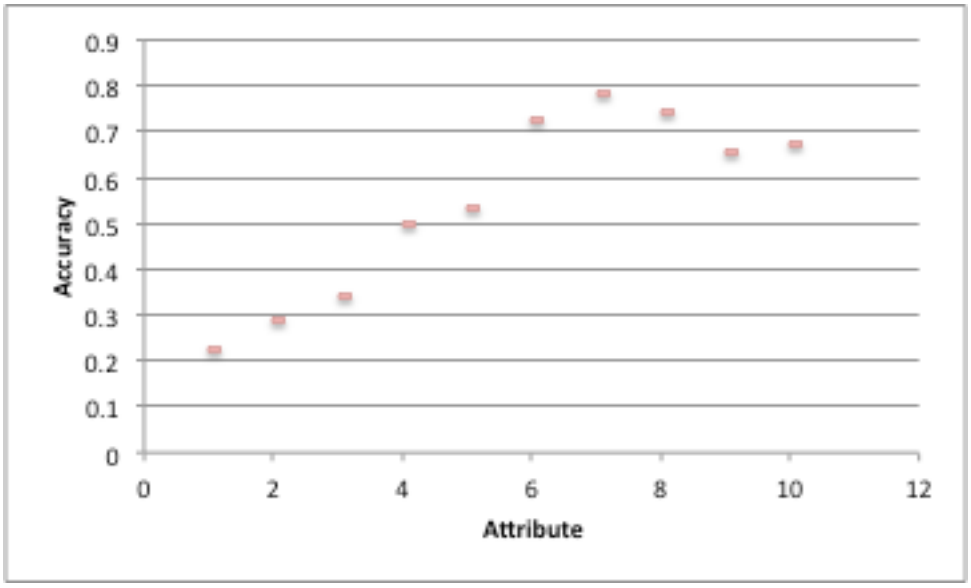

Figure 26: Accuracy Response by Attribute Type

The low reliability values for a small amount of attribute types are expected the amount of less time an identity is used, the higher the risk of the user if risk is dependent on reliability. Clearly, the attribute probabilities of compromise are used because they tend to level the accuracy of the risk metric. Again, it can be asserted that the relationship is not fully linear because Attributes 9 and 10 have a higher Probability of Compromise. The accuracy in this graph tends to follow the same pattern as that of the authenticity experiment. The risk accuracy is low for the first few sets of attributes but moves significantly higher then the authenticity graph after 6 attributes. This could make a case for combining the two metrics

5.5.2.2 Risk Precision Response to Variations in Number of Attribute Values per Type

During this experiment precision was also measured in risk values.

It was hypothesized that if the number of attributes values per identity is varied randomly while the number of transactions per identity remains constant, then the risk precision of each identity will increase as the number of attribute types per identity increases. The experiment ran with a variation of 1 to 10 attributes per identity for each transaction. The number of same values per attribute type was set at 20. Each identity conducted 50 transactions.

Again, the hypothesis was verified. The data collected is represented in Figure 27. In Figure 27, each diamond represents the precision for identity transactions with a certain number of attribute types. The precision is higher for those attribute types with larger numbers of similar values. Furthermore, when the amount of attributes per identity is low, the precision is also low.

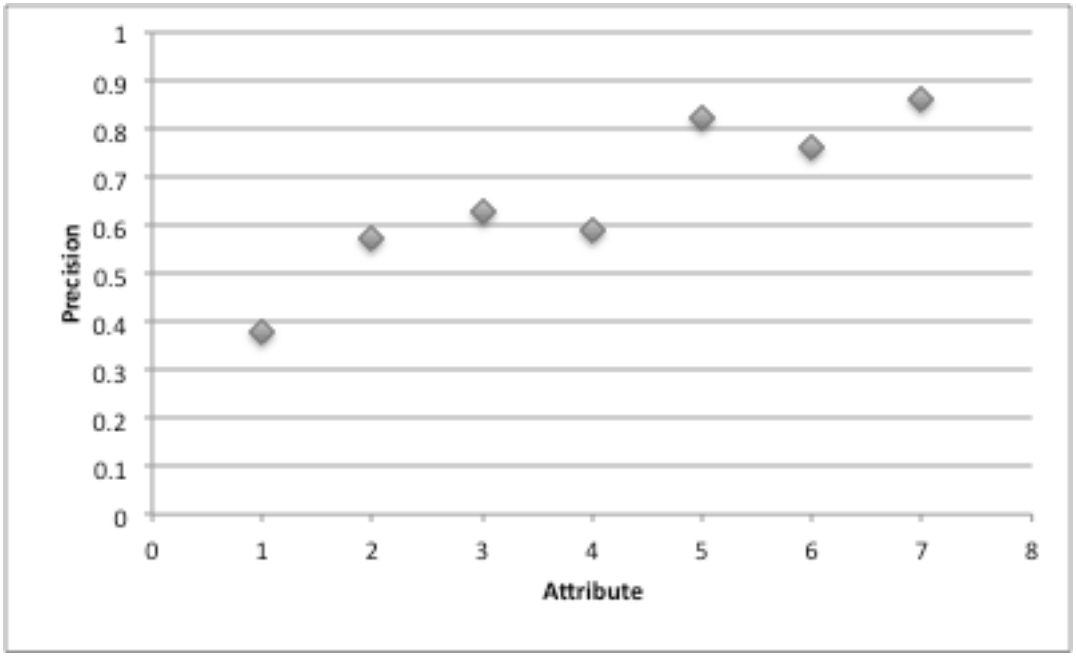

Figure 27: Precision Response by Attribute Type

The low precision values for small amount of attribute types are expected because an identity proves itself more authentic as it uses larger amounts of attribute types. These precision values for the risk metric using reliability are higher than those for authenticity.

5.5.3 Combination of Reliability and Authenticity to Determine Risk

In order to get a better risk values, the authors ran an experiment that combined authenticity and reliability. The risk metric was tested for variations in accuracy to changing numbers of attribute types. This experiment focused on all 5 useful measures of valid risk and was conducted in a series of transactions that took a set of 16 identities and varied the number of attribute types. Risk values were measured for each of these identities as each transaction was conducted following the same procedures in Chapter 4.

5.5.3.1 Risk Precision Response to Variations in Number of Attribute Values per Type

It was hypothesized that if the number of attribute values per identity are varied randomly while the number of transactions per identity remains constant then the risk precision of each identity will be increases as the number of attribute types per identity increases. The experiment ran with a variation of 1 to 10 attributes per identity for each transaction. The number of same values per attribute type was set at 20. Each identity conducted 50 transactions.

Again, the hypothesis was verified demonstrating changes in precision due to changes in values per type. The data collected is represented in Figure 28. In Figure 28, each blue diamond represents identity transactions with a certain number of attribute types. The reliability for each attribute is higher for those attribute types with larger

numbers of similar values. Furthermore, when the amount of attributes per identity is low, the reliability is also low. This is a useful indicator for calculating trust.

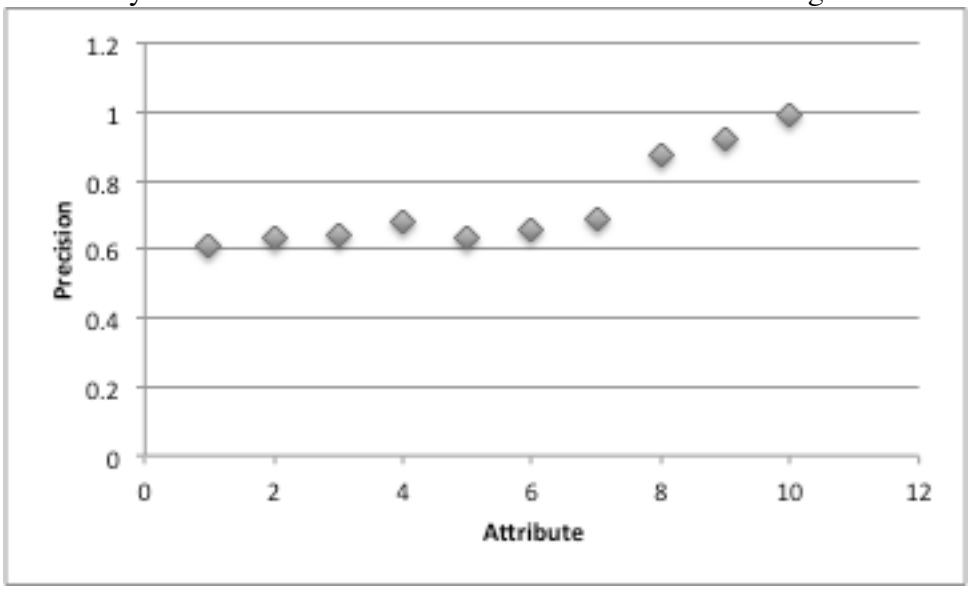

Figure 28: Precision Response by Attribute Type

The low precision values for a small amount of attribute types are expected because an identity proves itself more authentic as it uses larger amounts of attributes. These precision values for the risk metric using authenticity and reliability are higher than those for each metric individually.

5.5.3.2 Risk Accuracy Response to Variations in Number of Attribute Values per Type

It was hypothesized that if the number of attributes values per identity is varied randomly while the number of transactions per identity remains constant, then the risk accuracy of each identity will increase as the number of attribute types per identity increases. The experiment ran with a variation of 1 to 10 attributes per identity for each transaction. The number of same values per attribute type was set at 20. Each identity conducted 50 transactions.

Again, the hypothesis was verified demonstrating a significantly higher accuracy even at low attributes. The data collected is represented in Figure 29. In Figure 29, each data point represents mean accuracy for an attribute type. The accuracy clearly increases as the number of attribute types used increases. When the amount of attributes per identity is low, the authenticity and reliability are also low.

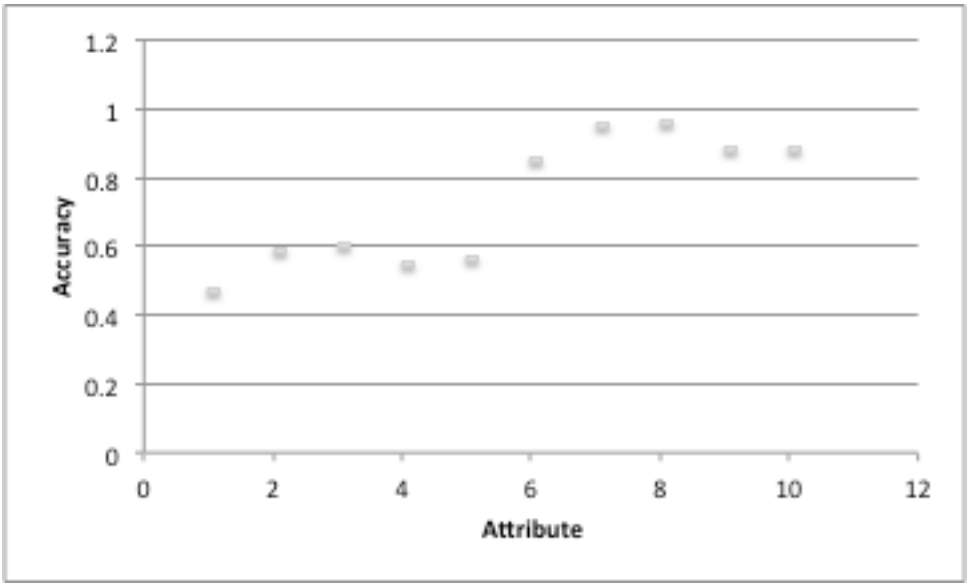

Figure 29: Accuracy Response by Attribute Type

When reliability and authenticity are combined, there is a much greater accuracy at each attribute type, which provides a much better predictor of overall risk. Service providers can leverage both of these metrics to determine an adequate level of risk calculation. Service providers' main focus has been to ensure that the right set of users is accessing it. If it is possible to accurately predict with even a small number of attributes used how risky a user is then it will make sense for the provider to adopt the methodology. Existing risk methods attempt to tie the risk to the type of data being accessed. By providing a risk view based on the user who attempts to access this data, the service

provider can flip that approach on its head. It can begin to adopt policies that use a certain level of risk to allow access to different domains. The types of information in the domains do not have to be formally assessed for consequences before a compromise probability can be applied to it. If the data is transactional in nature and meant to be accessed by users, then it is possible to this risk-based metric to meet service providers security needs.

Chapter 6: Analysis

Overall, this research attempts to leverage techniques based on trust to increase the confidence in an identity. Specifically the following hypothesis is considered:

In an online identity-based transaction ecosystem, trust in users and their attributes can be measured and used to enable less risky transactions.

Online transactions that leverage identity happen all of the time. Use of identity forms the bases for all online transactions that require trust. Users, service providers, and identity providers take unknown risks with every transaction involving an identity. Thus, the goal of this research is providing a means for reducing risk by quantifying it in these transactions. Once quantified, it is possible to use a risk value to make better judgments about how handling online transactions. In order to clearly verify this hypothesis, let us turn to several additional research questions.

6.1 Research Question 1

The definition of identity in Chapter One indicates that an identity can be computationally modeled by grouping its attributes together. With that understood, the first research question focuses on individual attributes:

RQ1: Which attributes of Identity are most trusted?

In order to answer the research question, the first necessary step is to define which attributes compose an identity. The related work section of this dissertation discussed a few different types of models [18, 85, 86, 88]. Eventually, the Identity Ecosystem was

chosen as an initial model because of the view that all attributes were connected. If all attributes of an identity are connected, then their connections must have some bearing on their utility. In fact, Reider asserts that attribute utility is proportional to its connectedness [9]. Each attribute model acknowledges that no provably complete identity models of attributes exist [85, 88]. A modeling approach was proposed to extend the identity attribute model as new attributes are discovered.

Often, it is asserted that some attributes can be similar or the same [85]. The modeling approach provides a way to differentiate through demonstration that certain attributes are different. Two attribute models are listed in Table 8 with the following comparisons. First, there are vastly different numbers of useful attributes with one having about 8% of the other. After comparing the attribute names, about 46% in the Army model had a corresponding value in the identity model. A conclusion to draw from the comparison is that differing identity providers had use for different types of attributes. This underscores the need for a reference model of attributes that functions across all identity providers.

Previous chapters described Probability of Fill and Probability of Compromise and measured the utility of the two metrics. It is also useful to compare them to existing research, but since identity security research is emerging, there are few significant studies for comparison. However, a few comments can be made. Probability of Fill is a good predictor of the usage of the attribute. If an attribute is filled at an identity provider, the identity user at some point must have used it. Thus this is a reliable predictor of usage but not a reliable predictor of its usage for any other function. Since Reiter asserts that effectiveness of attributes is proportional to their usage and their ability to be compromised [9], then the focus should be on they potential for compromise. The potential for compromise is related to how many places a user has used their attributes.

For example, if a user enters the same password across 10 different sites, then a compromise of any is a compromise of all. Just as important in Reiter's definition is the compromise potential, which is not just a function of how many places the attribute is used, but also how it could be derived from other attributes. If compromising one attribute can compromise another attribute, then the model must be able to account for all transitions between attributes. This attribute model showing the preferred attribute of the Ecosystem is able to satisfy that constraint.

It turns out that the method for calculation of Compromise of an attribute is also useful due to the correlation with recent studies on the most compromised attributes [32]. The correlation is expected because attackers exploit one attribute based on another. The correlation of the top 10 most compromised attributes to the choices made by the model are shown in the below Table 10.

Verizon Most Compromised Attributes	Measured most Compromised Attributes
Social Security Number	Social Security Number
Date of Birth	Date of Birth
Full Name	Mothers Maiden Name
Account Number	Email
Online Password	Account Number
Drivers License	Drivers License
Passport	Passport
Banking PINs	License Plate Number
Mothers Maiden Name	Medical Record Number

Table 10: Most Compromised Attributes

The proposed identity model employing the social model graph with attribute nodes and transition probabilities derived from ITAP found 60% of the top 10 attributes reported in the Verizon study with the most compromised being the Social Security

Number. This dissertation found 60% of the top 10 attributes as compared to the Verizon data without looking at the attributed usage data or Probability of Fill. If an attribute has a higher Probability of Compromise it is less trusted.

6.2 RESEARCH QUESTION 2

The definition of identity in Chapter One indicates that an identity can be computationally modeled by grouping its attributes together. Grouping these attributes together provides a computationally feasible identity for reasoning about trust.

RQ2: Can the reliability and authenticity of an identity in a transaction help demonstrate trust to an identity?

Trust is contextual. One minute an identity can be used in a correct manner by an authorized individual and the next the same identity or components of it could be stolen. Authenticity is primarily a function of trust in the contents of an identity used in a transaction. Therefore, the experiments generally reflected a strong correlation between the number of attributes used and the authenticity value. In order words, the trust is higher in a user who presents more attributes. Reliability is primarily a function of trust in the usage of identity in a transaction. Therefore an identity that uses attributes more consistently is shown as more reliable. Both measures of trust are necessary to accurately reflect the degree of trust placed in an identity. If just one or the other were used, it would be much easier to "game" the algorithm by exploiting its bias.

The experimental components of chapter 4 were designed to demonstrate that trust is useful in understanding the utility of multiple facets of an identity within a set of transactions. An identity is shown to be more reliable as both the numbers of transactions

increase and the numbers of attributes used for those transactions increase. This is a very useful result. For example, section 4.5.1 demonstrates that an identity is significantly more reliable when more then 4 attributes are used. This is a useful result to an identity provider seeking to design a protocol that used the least number of attributes that demonstrated the highest reliability.

Similarly, the number of transactions to gain a higher level of reliability is also described in the second experiment in 4.5.2. It took on average 26 transactions to pass the mean reliability threshold as shown in Table 11. These would be transactions in which a user presented fully valid credentials. This result has applicability to the protocol design community. Based on these results, it is possible to design a protocol that optimizes the number of samples of credentials to gain a certain threshold of reliability.

	Authenticity	Reliability
Mean level	0.464	0.4639
Average number of Transactions to reach mean	26 (1 Attribute) 34 (2 Attributes) 56 (3 Attributes) 73 (4 Attributes) 183 (5 Attributes)	26
Average number of attributes necessary to reach mean	6	6

Table 11: Authenticity and Reliability Data

The authenticity experiments also have useful results. The authenticity mean across these transactions is .4641. This is less important though then the slope of the authenticity curves. Each identity tested has a very sharp increase in authenticity at some point in the transaction set. Given the binomial nature of authenticity, it is useful as a

predictor because of the relative clustering of the data points at the low or high levels. RQ2 can be answered in the affirmative due to the responses in these metrics based on the representative set of data.

6.3 RESEARCH QUESTION 3

The work in previous chapters produced a computationally valid model for identity, which then can assume a value of trust in one of two dimensions: authenticity and reliability. Once the degree of trust in an identity is calculated, those values can be leveraged to calculate risk. The next research question reflects the nature of the risk calculated:

RQ3: Is the risk posed by an identity a reliable predictor of the validity of the identity?

In order to answer this question, it is necessary to use the trust metrics and then consider those as probabilities that an identity can be trusted. In the experiments, several sets of transaction were run with identities from the dataset described in RQ2. These transactions were then transformed into probabilities of risk. These probabilities of risk then were put into a Probit model to determine the thresholds that would be used to consider an identity risky (see Table 12).

The thresholds in the Probit model are telling because they indicate not only the level at which an identity will be considered risky but also a range of uncertainty. The combined risk model had the highest upper bound and the lowest percentage of uncertainty, with only 5 values or .625 percent of the data falling in between those bounds.

Probit Model	Risk (Authenticity)	Risk (Reliability)	Risk (Authenticity and Reliability)
Lower Bound	0.314	0.477	0.219
Upper Bound	0.686	0.523	0.780
Percent of data uncertain	2.88%	1.63%	.63%

Table 12: Probit Model Thresholds

According to the Probit model, the combined authenticity/reliability risk metric provides the highest resolution on the data with the authenticity, and reliability not far behind. A graphical view of how the data spreads out can be seen in Figure 30 for the authenticity risk metric. Overall, the sensitivity of the model demonstrates clearly that the modeling approach provides a sound way to make a judgment of whether an identity is risky or not.

Given the ability of the model to make a judgment of riskiness, the next level is the ability for the model to be an accurate judge of that risk. Table 13 compares different types of risk calculated and the improved reliability metric. The improved reliability metric can be used as a guide for determining whether an identity is correct.

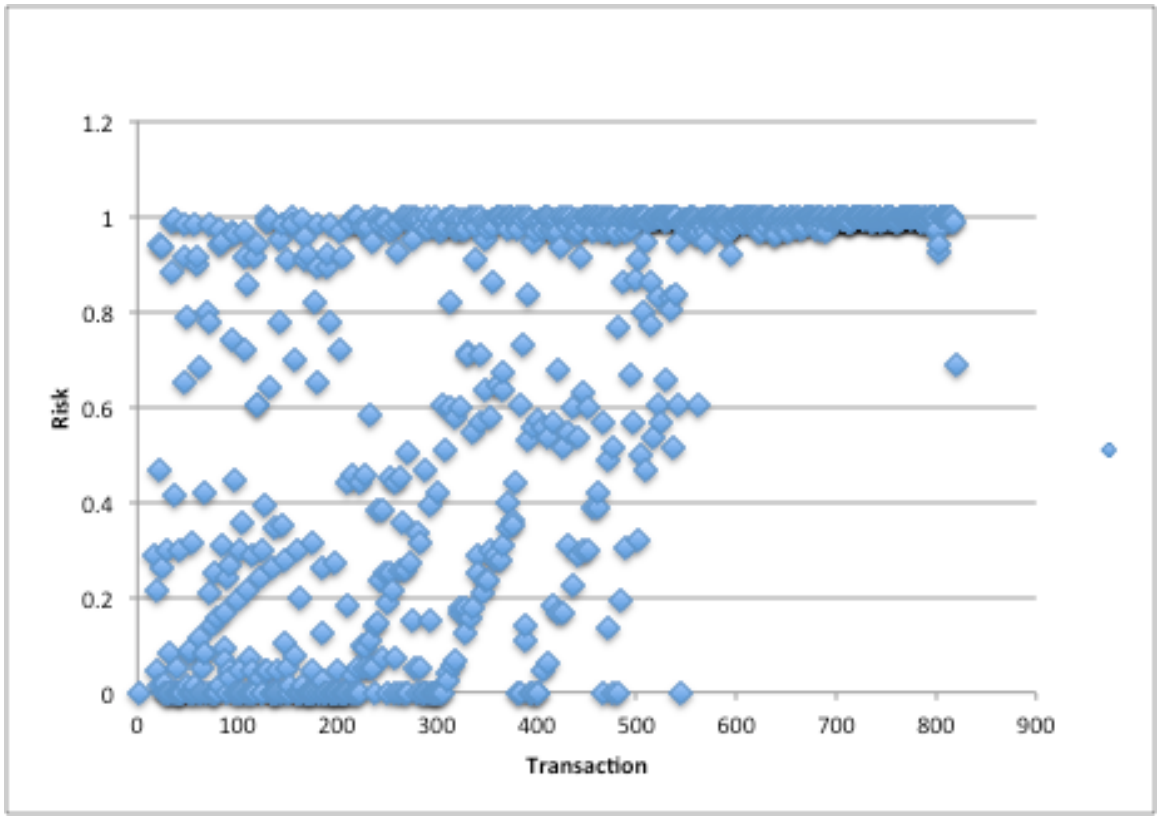

Figure 30: Authenticity Risk vs. Transactions Conducted

	Risk (Authenticity)	Risk (Reliability)	Risk (Authenticity and Reliability)	Improved Reliability
Mean Average Precision	0.553	0.739	0.733	0.564
Average Accuracy	0.567	0.544	0.727	0.269
Rank	2	3	1	4

Table 13: Rankings of Different Types of Risk Metrics

These metrics were each run against the same set of transactions to determine which had the higher accuracy in prediction of a true identity. Rather then comparing risk values, the best comparison to determine effectiveness is accuracy of the metric in determining

whether an identity matches one inside the golden set. The model is configured such that if the risk value is higher then the Probit upper bound then it is considered a valid identity. If the risk value is lower then the Probit lower bound then it is considered an invalid identity. The combined authenticity/reliability risk calculation turns out to be the best predictor out of the 4 different risk calculations.

The research question can be answered in the affirmative that risk is a good predictor of identity to the degree that it is 72.6% accurate.

Chapter 7: Conclusion

Identification of a person or a thing remains a difficult problem but not intractable. This work will form the basis for a new approach to online identification. While identity security is often about protecting the Personal Identifiable Information (PII) with techniques such as encryption and network, the heart of identity trustworthiness is not about securing the PII but determining if the PII can be trusted. This dissertation has taken a probabilistic view of identity asserting that there is no absolute identification. Those probabilities have been evaluated across common Internet transactions such as authentication and authorization.

Many techniques and methodologies have been used in research to establish identity. Most of these processes involve some form of strengthening of the base identity with additional attributes such as biometrics. Much information already exists about people on the Internet, and identity providers capture it. This existing information can strengthen the validity of identities used. The lack of focus on the model of identity also drives a heavy reliance on increasing security on each transaction by developing new techniques and processes for needs such as authentication. This dissertation provides a supplementary or alternate method to leverage existing information stored and transmitted during transactions to increase confidence in the user's identity without requiring additional information.

Leveraging PII models from the Center for Identity [18] and the U.S. Army [101], this research reviewed identity models and the information used to comprise those models. The dissertation results assert that the connectedness of attributes in identity is a good predictor of their compromise. Typical approaches are limited by relying on information that spreads across multiple physical locations of identity in order to

understand which attributes can be compromised and to what level of probability. This dissertation looks across identity providers by simply looking at the personal model of identity attributes.

In order to increase confidence through the usage of additional attributes, a model had to be developed that describes identity by its attributes. A number of researchers have attempted to generate a standard taxonomy or model around identity. None can adequately represent the identity as a computationally feasible model of attributes. The attribute model used inside this dissertation will enable security researchers to generate more robust models and perform predictive analysis. The modeling process defined will enable additional attributes to be added and analyzed. Further, the addition of the attributes to compose a system identity model provide the ability to reason about and understand the behavior of other identity systems which are under constant attack today. In fact Identity Providers will be able to use those attributes at the time of identification and throughout the transaction to inform which attributes are requested. Identity Providers will also be able to choose the correct set of attributes needed to inform their particular needs. In the long run, this can slow the significant growth in the amount of attributes being requested of users.

This research developed a trust model directly correlated to the attributes in identity to establish a degree of trustworthiness. Leveraging identities and transactions developed using techniques by the Department of Homeland Security Identity resolution team [92] the trust model was evaluated. Trust was decomposed into authenticity and reliability with each providing different information about how the identity could be trusted. The reliability measure was also extended to determine the degree to which the Probability of Compromise of certain attributes affects reliability. Trust as a notion is heavily used in a number of different areas such as modeling agents behavior [3], and in

peer to peer transactions [5]. It has never been applied to an abstract notion of identity. This dissertation shows that identity has sufficient properties to compute trust models between an identity and identity provider. These can be used transitively to provide access to services. The first trust metric proposed is an identity-based reliability metric that providing a concrete confidence value rather then assuming implicit trust in the system. This reliability metric can be used to contribute to decisions in Authentication schemes for people and system based services. Another contribution is the calculation of the authenticity of an Identity. Although many reputation generators exist for trust-based approaches, adding authenticity as a property to Identity allows it to become more durable and usable in computations. This measure was called improved reliability. Improved reliability was found to be a less accurate measure of a compromised identity then risk as investigated in RQ3.

Finally, a risk model offered by this dissertation research leverages the following calculations to ultimately determine whether risk is a good predictor of an invalid identity.

- Reliability – The component of trust that expresses whether an identity is used correctly repeatedly. Reliability provides for partial valid usage.
- Authenticity – This component of trust expresses whether an identity uses the correct attributes. It also provides a view of uniqueness of those attributes
- Probability of Compromise – The probability that an attacker will be able to compromise any specific identity attribute given a connected attribute is known.

The measurements were combined using a Bayesian network built around the identity model to determine a risk probability. Using the three together will make it harder for attackers to exploit and game the system to provide favorable risk values for exploitation. This risk probability was then divided using a Probit model to determine

whether the risk leaned toward a valid or invalid identity. The risk probability will also supplement service providers' ability to grant access to resources over time. Based on their needs, service providers will be able to dial up or down risk to access their resources. This approach allows for an overall more secure approach to providing services to end users taking into account existing vulnerabilities on the Internet. The risk number turned out to be a useful predictor of whether an identity was valid.

7.1 FINDINGS

There are three main findings that can be drawn from this research. First, **an identity attribute's chances of compromise are strongly related to connectivity the attribute to other attributes.** The experiments related to Research Question 1*:Which attributes of Identity are most trusted?* demonstrated that by using social graph preference techniques, empirical data can be represented about how identity attributes are compromised using an attribute to acquire other attributes. From these experimental results, it can be determined that the technique applied to assess the Probability of Compromise and the top most compromised attributes is useful in assessing how data is breached and the identity attributes most likely to be breached. In comparison to the Verizon 2013 data breach report [48], this dissertation research agreed with 6 of the 10 most compromised attributes identified by Verizon. The attribute analysis approach was also determined to be repeatable and can evolve as the understanding of identity thieves evolves in the Center for Identity Threat Assessment and Prediction (ITAP) data set. Also, identity providers could use these techniques to more effectively collect identity attributes for authentication and more effectively secure the identity attributes they collect.

Second, **trust is a good predictor of identity.** The experiments related to Research Question 2: *Can the reliability and authenticity of an identity in a transaction help demonstrate trust to an identity?* demonstrated that by applying authenticity and reliability measures to sets of identity data, it is possible to quantify the trust in a particular identity based on its attributes. An identity is determined to be more authentic as it presents more correct attributes. The research also substantiates that reliability is a good predictor of identity based on the number of similar transactions conducted. An identity that has both high authenticity and reliability tends to be most likely to be trusted and correct. From these experimental results it can be determined that authenticity and reliability are good predictors of trust. The probability of a user being authentic correlated directly with number of valid users in the experiments. The probability of a user being reliable correlated directly with the number of valid users in the experiments. Neither authenticity nor reliability however by themselves provided a strong correlation to the number of valid users. Thus risk was introduced in the next research question to overcome those limitations. Further experimental results demonstrated that the degree of authenticity and the degree of reliability were both positively correlated with the amount of attributes used during transactions. They were also positively correlated with the amount of transactions conducted. Practically an identity provider could apply these findings by requesting more attributes to gain a higher level of trust in a user. Today some identity providers require additional information to more positively identify a user but there is no quantifiable model as to how much information to collect about a user. Also, this research finds that requesting more transactions can also drive a higher level of trust in a user. Increasing the number of transactions is not typically applied in the identity space as a method to increase trust. Identity providers could establish

mechanisms that forced a certain level of transactions in order to gain a higher degree of trust in their end users.

Third, this dissertation concludes that **risk is a strong predictor of identity compromise.** The experiments related to Research Question 3*: Is the risk posed by an identity a reliable predictor of the validity of the identity?* Demonstrated that each component of trust by itself was positively correlated with risk. Those positively correlated components included: Authenticity, Reliability, Compromise and Improved Reliability. When combined with each other and the probability that an attribute is compromised, the prediction accuracy significantly increased. From these experimental results it can be determined that risk can predict 72% of valid and invalid identities for a given set of transactions using experimental data. This result is significant because the existing mechanisms would assume each of these identities was valid based solely on their authorization thus grant the user (the identity) access. Adding a risk identification methodology would give service providers greater confidence about how risky the users are that access their data. Further, the experimental results demonstrate no need for the definition of consequence. Existing risk methodologies explored required some knowledge of the data being accessed in order to define the risk to that data. This finding turns that on its head by choosing the riskiness of the user accessing the data. The flexibility gained allows service providers to use a identity-centric risk metric to determine data security by not giving risky users access in the first place and; thus, not having to determine the very difficult estimation of the consequence of data release.

7.2 LIMITATIONS

Although this research proves the research questions posed, there are limitations to how far reaching the conclusions are to be used. The first limitation is the extent to

which the dataset can be interpreted. There are approximately 350M people in the US as of 2014. The dataset sampled 50K of these people yielding a representative sample of .014%. Statistically this sample size yields a confidence interval of .25 with a 95% confidence level. Although the work is considered statistically significant, the confidence level could increase with a larger dataset.

Another limitation is the lack of real transaction data. As discussed in RQ2, the transaction data was structured following the patterns used by US Department of Homeland Security for testing of its systems. While it is unclear what assumptions made in the construction of this data, it is possible that some techniques used might introduce bias. While Miller, et al.[92] was very clear about the source of data and the composition of it at a macro level(as mentioned above) he didn't describe certain aspects. One was how he chose the record set from the 77 million decedents inside the master data file. Was it a random sampling, if so how were the cultural constraints met? This potential source of bias could lead to large amounts of people with same names or attributes that aren't representative of the population. The authenticity metric is dependent on those numbers. For example, if an unusually large amount of Brian's were chosen and Brian was used as an attribute in a transaction, that might bias the authenticity for that Identity to be lower. Any other limitations that exist within this work are specifically discussed in the analysis of the experiments or data.

7.3 Future Work

It is possible to extend this work in the future to develop more accurate and precise models of identification. The first and foremost area that deserves more research is the categorization and modeling of attributes. While researching RQ1, it became clear that almost anything could be an attribute as long as it fits the criteria listed. Thus

relationships can grow large and complex. The modeling strategy though, constrains this model to a tractable set of attributes. These attributes can be found, recorded, and modeled. The larger this model grows, the more precise the probabilities for compromise will be. The Center for Identity at the University of Texas has a program called ITAP that continually mines new attributes and relates them to each other in a model.

Another research opportunity is to consider attribute values more deeply. This research only looks for direct matches of attributes and it does not account for probable matches. For example, there are multiple spellings or nicknames of Brian- Bryan, Bri, Bry, etc. These potentially have the same meaning and refer to the same identity record in the identity store. Semantic ontologies have the ability to find probabilistic matches of these attribute values. These probabilities could build on top of the existing trust and risk algorithms to increase accuracy of the results without increasing the likelihood of false positives.

Finally, another area of research in business risk could be employed. Many risk models incorporate the dimension of Impact of an event. For example, it is possible that the risk values go up for certain identities based on who they are, how connected they are, or what types of access they have. These are all important factors to consider as an Identity provider who uses identity for many different types of transactions. Introductions of this dimension could make the risk calculation a useful commercial tool for businesses.

Transactions between users on the Internet require credentials that have a fixed number of attributes. When these credentials are created, attributes such as Social Security number, mother's maiden name, and address are used to validate them. Attributes are often lost, stolen or compromised. Once the attributes of an identity are compromised, anyone can assume that identity with benign or malicious purposes.

Traditional solutions to this problem are to increase the trust level of the authentication through multiple modes, such as biometrics or smartcard tokens. This research shows that it is possible to increase trust of users without requiring these extra items. Using only the attributes registered with an identity provider (e.g., address, zip code, name, etc.) It is possible to show how trusted a user is who presents an identity. Further, the risk to a service provider of allowing access to that user can be established with this limited information.

.

References

1. D. Gefen, Reflections on the dimensions of trust and trustworthiness among online consumers, ACM SIGMIS Database 33 (3) (2002) 38–53.
2. D.W. Manchala, E-commerce trust metrics and models, IEEE Internet Computing 4 (2) (2000) 36–44
3. K Fullam, Adaptive Trust Modeling in Multi-Agent Systems: Utilizing Experience and Reputation, UT Dissertation, 2007
4. S Marsh, Formalising Trust as a Computational Concept, Dissertation1994
5. L Xiong, L Liu. Building Trust in Decentralized Peer-to-Peer Electronic Communities, Building Trust in Decentralized Peer-to-Peer Electronic Communities, 2005
6. Y. Liu, "Trust-Based Access Control for Collaborative System". Proc. CCCM 2008, IEEE Computer Society, pp. 444-448, 2008
7. T. Yu, S. Hartman, K. Raeburn RFC 4120, MIT July 2005
8. Various, Authentication Context for the OASIS Security Assertion Markup Language (SAML) V2.0, 2005
9. K. Reiter and S. G. Stubblebine, Toward Acceptable Metrics of Authentication, Proceedings of the IEEE Symposium on Security and Privacy, pp.10-20, Oakland, May 1997.
10. R. S. Sandhu, E. J. Coyne, H. L. Feinstein, and C. E.Youman, "Role- based access control models", IEEE Trans. Computer, vol.29, no.2, pp. 38- 47, 1996.
11. L. Snyder, "Formal models of capability-based protection systems", IEEE Trans. Computer, vol.30, no.3, pp. 172-181, 1981
12. H. Wen, L. Fang, L. Guan;A MULTI-AGENT BASED AUTOMATIC WEB RECOMMENDATION MODEL, Proceedings of the Eigth International Conference on Machine Learning and Cybernetics. 2009
13. J. Kilian and E. Petrank. Identity Escrow. In CRYPTO 1998, Springer-Verlag (LNCS 1462), pages 169–185, 1998.
14. KS Barber, J Kim Belief Revision Process based on Trust: Agents Evaluating Reputation of Information Sources, 2002.
15. S. E. Schechter, T. Parnell and A.J. Hartemink. Anonymous Authentication of Membership in Dynamic Groups. In 3rd Financial Cryptography, Springer-Verlag (LNCS 1648), pages 184–195, 1999.

16. A Lindell. Anonymous Authentication, 2005

17. S Agrawal, Efficient, Provably Secure Code Constructions, University of Texas Disseration, 2011.

18. KS Barber, S Budalakoti. A Bayesian Network based Framework for Identity Risk Management, Technical Report- TR-UTCID 13042013, The University of Texas at Austin, 2013

19. P Kearney, L Brugger, A Risk Driven Security Analysis method and modeling language, BT Technology Journal Volume 25, No 1 January 2007

20. Maurer, Ueli Modelling a public-key infrastructure, Computer Security—ESORICS 96, 1996

21. B. Clifford Neuman and Theodore Ts'o (September 1994). "Kerberos: An Authentication Service for Computer Networks". IEEE Communications 32 (9): 33–8

22. B.A. Gran, R. Fredriksen, and A.P.-J. Thunem, An Approach for Model-Based Risk Assessment, In Proceedings of the 23rd International Conference on Computer Safety, Reliability, and Security (SAFECOMP 2004),volume LNCS 3219,Springer-Verlag, pp. 311-324,Potsdam,Germany,September,2004.

23. M Blaze, J Feigenbaum, J Lacy. Decentralized Trust Management, IEEE Symposium on Security and Privacy pg 164-173, 1996

24. Kini A. and Choobineh J. Trust in Electronic Commerce: Definition and Theoretical Considerations, in 31st Annual Hawaii International Conference on System Sciences, 1998, Hawaii

25. Lorch, M.; Basney, J.; Kafura, D.; , "A hardware-secured credential repository for Grid PKIs," *Cluster Computing and the Grid, 2004. CCGrid 2004. IEEE International Symposium on* , vol., no., pp. 640- 647, 19-22 April 2004Martin, Raymond and Baressi, John Personal Identity, Blackwell Publishing Ltd, 2003

26. Bolme, David, Beveridge, J. Ross, and Howe, Adele; Personal Identification Using Text and Image Data, 2007

27. P. Verlinde, G. Chollet, M. Acheroy, Multi-modal identity verification using expert fusion, Information Fusion 1 (2000) 17– 33.

28. Bengio, S., Marcel, C., Marcel, S., & Mariéthoz, J. (2002). Confidence measures for multimodal identity verification. *Information Fusion*, *3*(4), 267–276. doi:10.1016/S1566-2535(02)00089-1

29. J Kittler, J. Matas 1 K Jonsson M U Ramos Sanchez, Combining evidence in personal identity verification systems

30. Federal Identity Credential and Access Management Roadmap and Implementation Guidance, November 10, 2009

31. Sanderson, C., & Paliwal, K. K. (2004). Identity verification using speech and face information. Digital Signal Processing, 14(5), 449–480. doi:10.1016/j.dsp.2004.05.001

32. Javelin Strategy and Research- 2012 Identity Fraud Report- Consumers taking control to reduce their risk for fraud

33. H. Adkins. An update on attempted man-in-the-middle at- tacks. Technical report, Google Online Security Blog, Aug. 2011. http://googleonlinesecurity.blogspot.com/2011/08/ update-on-attempted-man-in-middle.html.

34. C Metz AAA protocols: authentication, authorization, and accounting for the Internet. Internet Computing IEEE Nov/DEC1999 pg 75-79

35. W. Stallings, Protect Your Privacy, A Guide for PGP Users, Prentice Hall 1995.

36. T. Dimitrakos, B. Ritchie, D Raptis, K Stolen Model based Security Risk Analysis for web applications: The Coras Approach; Euroweb 2002

37. M Kohli, Tranformation from Identity Stone Age to Digital Identity, International Journal of Network Security and Its applications, May 2011 pg 121-136

38. C.E. Bonafede, P. Giudici, Bayesian networks for enterprise risk assessment, JUL 2006

39. P Bissiri, S. Walker, On Bayesian Learning from Bernoulli Observations, JUN 2010

40. J Huang, D Nicol, A Calculus of Trust and Its Application to PKI and Identity Management.

41. Li XY, Gui XL,A Comprehensive and Adaptive Trust Model for Large- Scale P2P Networks. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY,2009,24(5): 868-882

42. National Strategy for Trusted Identities in Cyberspace- Presidential Directive April 2011

43. The Presidents Identity Theft Task Force Report- Committee, September 2008

44. Federal Trade Commission, Consumer Sentinel Data Book March 2011

45. D Forencio, C Herley, A Large Scale Study of Password Habits, WWW 2007, May 2007

46. M Bertin, Market Trends and Forecasts- World Card Summit, November 15, 2011

47. Pulled from www.internetworldstats.com/stats.htm on June 18 2012

48. Verizon Data Breach report 2013

49. T Perrin, L Bruns, J Moreh, T Olkin, Delegated Cryptography, Online Trusted Third Parties, and PKI, 1st Annual PKI Research Workshop, April 2002

50. Ferguson, Niels; Bruce Schneier (2003). *Practical Cryptography*. John Wiley & Sons. ISBN 0-471-22357-3.

51. Data Retrieved from Alexa.com on June 19, 2012

52. Data Retrieved from Alexa.com on June 19, 2012

53. E Yuan, J Tong, Attribute Based Access Control for Web Services, 2007

54. D Ferraiolo, R Sandhu, S Gavrila, D Kuhn, R Chandramouli, Proposed NIST Standard for Role-Based Access Control, ACM Transactions on Information and System Security August 2001

55. Josang, Audun and Ismail, Roslan and Boyd, Colin A. (2007) A survey of trust and reputation systems for online service provision. Decision Support Systems 43(2):pp. 618-644.

56. RFC4520

57. R Chisholm, Parts as Essential to the Wholes, The Review of Metaphysics 1973, vol 26(4) pg 581-603

58. T Beth, M Borcherding, and B Klein. Valuation of Trust in Open Networks, Proceedings of the European Symposium on Research in Computer Security, 1994.

59. Reagle Jr., J. M. 1996. Trust in a cryptographic economy and digital security deposits: Protocols and policies. Master of Science Thesis, Department of Technology and Policy, MIT.

60. J Treur, C. M. Jonker. Formal Analysis of Models for the Dynamics of Trust based on Experiences. 1999

61. DJ Kim, DL Ferrin, HR Rao. A trust-based consumer decision-making model in electronic commerce: The role of trust, perceived risk, and their antecedents, Decisions Support Systems Conference 44, 2008, pg 544-564

62. P Zimmerman. PGP Users Guide, Volumes I and II. October 1994

63. L. Krautsevich, A. Lazourski, F. Martinelli, A. Yautsiukhin, Risk-aware Usage Decision Making in Highly Dynamic Systems, Fifth International Conference on Internet Monitoring and Protection, 2010

64. D Kahneman and A. Tversky. Prospect Theory An Analysis of Decision under Risk, Econometrica 1979.

65. Martin, Raymond and Baressi, John Personal Identity, Blackwell Publishing Ltd, 2003

66. Retrieve from Google Analytics 20AUG 2012 https://developers.google.com/analytics/devguides/collection/gajs/gaTrackingEcommerce
67. Retrieved from Click Tracks Website- 20 AUG 2012 http://support.clicktracks.com/
68. B. Krishanmurthy, CE. Wills. On the leakage of Personally Identifiable Information via online social networks. Internaltional World Wide Web Conference Committee, WWW 2009.
69. B. Malin, L. Sweeney, and E. Newton. Trail re-identification: learning who you are from where you have been. LIDAP-WP12. Carnegie
70. Mellon University, Laboratory for International Data Privacy, Pittsburgh, PA: March 2003.
71. B Liu, H Lu, Y Zhao, F Ge, A Framework Trust Establishment for E-Services, 2010 International Conference on e-Education, e-Business, e- Management, and e-Learning
72. R Falcone, G Pezzulo, C Castelfranchi, A Fuzzy Approach to a belief based trust computation. 2003
73. C Basu, H Hirsh, W Cohen, Recommendation as Classification: Using Socal and Content Based Information in Recommendation. Proceedings of AAAI 1998.
74. R. Haenni, J. Jonczy, R. Kohlas. Two Layer models for Managing Distributed Authenticity and Trust. 2007
75. J Muncaster, M Turk. Continuous Multimodal Authentication using Dynamic Bayesian Networks, UCSB 2005.
76. J. Bigham, D. Gamez, N. Lu. Safeguarding SCADA Systems with Anomaly Detection, Computer Network Security 2003
77. VM Igure, SA Laugher, RD Williams Security issues in SCADA networks, Computers and Security, 2006
78. RE Dawson, C Boyd, E Dawson, JM Gonzalez Nieto, SKMA - A Key Management Architecture for SCADA systems, Austrailian International Information Security Workshop, 2006
79. AA Cardenas, T Roosta, S Sastry, Rethinking security properties, threat models, and the design space in sensor Networks: A case study in SCADA systems, AD HOC Networks, 2009
80. A Thabet, Stuxnet Malware Analysis Paper, Retrieved from http://www.codeproject.com/KB/web-security/StuxnetMalware/Stuxnet_Malware_Analysis_Paper.pdf

81. M Rahbari, MA Jabreil Jamali, EFFICIENT DETECTION OF SYBIL ATTACK BASED ON CRYPTOGRAPHY IN VANET, International Journal of Network Security and its Applications, NOV 2011

82. JR Douceur, The Sybil Attack, Proceedings of 1st International Workshop on Peer to Peer Systems, 2002

83. M Ion, L Telesca, F Botto, H Koshutasnski. An Open Distributed Identity and Trust Management Approach for Digital Community Ecosystems. 2008

84. A Wagner, T Dubendorfer, B Plattner, R Hiestand, Experiences with Worm Propagation Simulations, Worm '03, OCT 2003

85. Thomas, I., & Meinel, C. An attribute assurance framework to define and match trust in identity attributes. In *Web Services (ICWS), 2011 IEEE International Conference on* (pp. 580-587). IEEE.

86. Mueller, Milton L., et al. "Digital identity: How users value the attributes of online identifiers." *Information Economics and Policy* 18.4 (2006): 405-422.

87. Ahn, Gail-Joon, and John Lam. "Managing privacy preferences for federated identity management." *Proceedings of the 2005 workshop on Digital identity management*. ACM, 2005.

88. Multiipel, NASPO IDPV Project *Establishment of Core Identity Attribute Sets and Supplemental Identity Attributes,* Document No. NASPO-IDPV-060, Feb 2014

89. Holton, Glyn A. "Defining risk." *Financial Analysts Journal* (2004): 19-25.

90. Marforio, Claudio, Nikolaos Karapanos, and Claudio Soriente. "Smartphones as Practical and Secure Location Verification Tokens for Payments." (2014).

91. L. Jean Camp. "Digital identity." *Technology and Society Magazine, IEEE* 23.3 (2004): 34-41.

92. Keith J. Miller, Elizabeth Schroeder, Sarah McLeod, Azar Ulrich, Karine Megerdoomian, James Finley, Gail Hamilton, Andre Milota, Ken Samuel, Sherri Condon and Mark Arehart. 2009. Method for comparing identity resolution technology. Presentation at *77th MORSS*, Fortrei Leavenworth, Kansas

93. Dirk Neumann, Mark Baker, Jorn Altmann, and Omer F. Rana. 2009. *Economic models and algorithms for distributed systems*. Springer, Dordrecht, Germany.

94. Jeff Jonas. 2006. Identity resolution: 23 years of practical experience and observations at scale. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data* (SIGMOD '06). ACM Press, New York, NY, 718-718. DOI=10.1145/1142473.1142556 http://doi.acm.org/10.1145/1142473.1142556

95. Brian Soeder and Kathleen S. Barber. 2014a. Towards a metric for confidence in identity: An agent based approach. In *Proceedings of the International Conference on Agents and Artificial Intelligence*.

96. Brian Soeder and Kathleen S. Barber. 2014b. Trustworthiness of identity attributes. In *Proceedings of the 7th Annual Conference on Security of Information and Networks* (*SINCONF 14*).
97. Mohsen Jamali and Martin Ester. 2009. *TrustWalker*: a random walk model for combining trust-based and item-based recommendation. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '09). ACM, New York, NY, USA, 397-406. DOI=10.1145/1557019.1557067 http://doi.acm.org/10.1145/1557019.1557067
98. Page, Lawrence, et al. "The PageRank citation ranking: Bringing order to the web." (1999).
99. Kim, Dan J., Donald L. Ferrin, and H. Raghav Rao. "A trust-based consumer decision-making model in electronic commerce: The role of trust, perceived risk, and their antecedents." *Decision support systems* 44.2 (2008): 544-564.
100. Wasserman, Stanley. *Social network analysis: Methods and applications*. Vol. 8. Cambridge university press, 1994.
101. Army Data set obtained from Army Knowledge Online SEP 2013.
102. Interview with Identity experts, NOV 2013
103.Yongpeng Yang; Manoharan, M.; Barber, K.S., "Modelling and Analysis of Identity Threat Behaviors through Text Mining of Identity Theft Stories," Intelligence and Security Informatics Conference (JISIC), 2014 IEEE Joint , vol., no., pp.184,191, 24-26 Sept. 2014

Vita

Brian Soeder was born in Niagara Falls, NY. He attended Amherst High School in Amherst, NY. Afterward, he was awarded an Army ROTC scholarship and attended in Rensselaer Polytechnic Institute in Troy, NY. After graduation with a BS in Computer and Systems Engineering, he entered the US Army. Brian was assigned to the Signal Corps and performed in various leadership capacities at Fort Hood, TX including command of a Field Artillery Battery. After he completed service in the Army, he took an assignment to Kuwait to support the war effort in Operation Iraqi Freedom where he was the Senior Engineer in theatre supporting collaboration between coalition forces. Next, he began a job with The MITRE Corporation performing in various capacities including Member of the Technical Staff, Project Lead, and Technical Lead. While at MITRE he obtained his MS in Software Engineering from the University of Texas. He is currently serving as the Group Lead for the Software Engineering Group in MITRE's Army Division.

3014 Raindance Loop, Harker Heights, TX 76548

This dissertation was typed by the author.